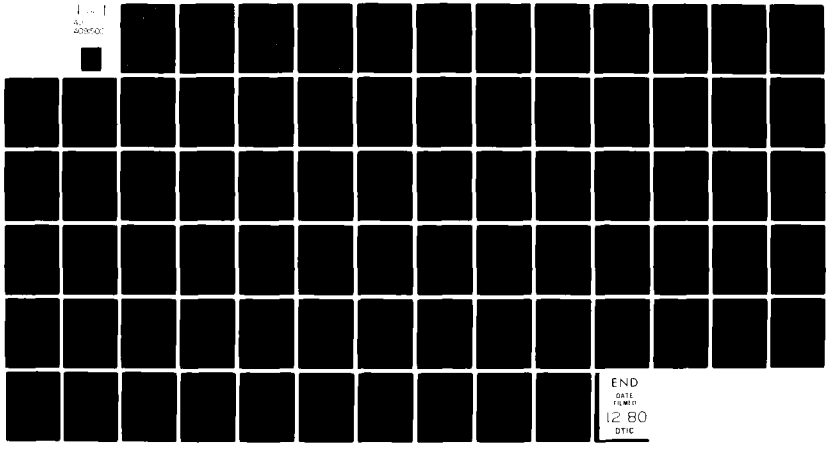
AD-A091 502 ILLINOIS UNIV AT URBANA-CHAMPAIGN DEPT OF COMPUTER SCIENCE F/G 12/1
NUMERICAL METHODS FOR INITIAL VALUE PROBLEMS.(U)
JUL 80 R D SKEEL AFOSR-75-2854

UNCLASSIFIED AFOSR-TR-80-0985 NL



END
DATE FILMED
12 80
DTIC

UNCLASSIFIED

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO.	3. RECIPIENT'S CATALOG NUMBER
AFOSR-TR-80-0985	AD-A091 502	

4. TITLE (and Subtitle)

NUMERICAL METHODS FOR INITIAL VALUE PROBLEMS

5. TYPE OF REPORT & PERIOD COVERED: Rept. Final 6/1/75-6/30/80

6. PERFORMING ORG. REPORT NUMBER: 1 Jun 75-30 Jun 80

7. AUTHOR(s)

Robert D. Skeel

8. CONTRACT OR GRANT NUMBER(s)

AFOSR-75-2854

LEVEL

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Department of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

61102F
2304/A3
A3

11. CONTROLLING OFFICE NAME AND ADDRESS

Air Force Office of Scientific Research/NM
Bolling AFB, DC 20332

12. REPORT DATE: July 1980

13. NUMBER OF PAGES: 77

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

78

15. SECURITY CLASS. (of this report): unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

AD A091502

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited

DTIC ELECTE NOV 7 1980 C

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

error estimation, deferred correction, numerical solution of hyperbolic PDEs, Gaussian elimination, multistep methods, multigrid methods

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Computational techniques were studied for the estimation of global discretization error in numerical solutions of both smooth and nonsmooth differential equations with an emphasis on the deferred correction technique. Of special interest were hyperbolic problems, for which mixed results were obtained. The implications of a stronger stability concept for Gaussian elimination were explored with respect to scaling, iterative refinement, and equilibration. Interesting equivalence and stability results were obtained for multistep methods for solving ordinary differential equations. The rapid convergence

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

DDC FILE COPY

176 011

20. of a multi-level algorithm for elliptic problems on locally refined grids was demonstrated.

NUMERICAL METHODS FOR INITIAL VALUE PROBLEMS

2 July 1980

AFOSR-TR- 80-0985

RESEARCH OBJECTIVES

FINAL

(3)

The primary objective of this work is to determine computational techniques for estimating the error present in the solution of time-dependent partial differential equations for both smooth and nonsmooth problems and in particular for hyperbolic problems with shocks. Another objective is to study variable step versions of general multistep methods for ordinary differential equations and to implement an efficient algorithm for the solution of stiff equations. Still another objective involves the study of the multi-grid method for solving linear algebraic systems arising from discretized elliptic partial differential equations.

AFOSR 75-2854

STATUS OF THE RESEARCH EFFORT

The research reported here is in the area of numerical analysis. There are four parts to this report:

I. Computational error estimates and deferred corrections for differential and integral equations

II. Roundoff error for variants of Gaussian elimination

III. Multistep methods for ordinary differential equations

IV. Multi-grid methods for elliptic partial differential equations

A more detailed outline is given by the table of contents, which follows.

The work described in this report falls into three categories: (i) work that is reported elsewhere, for which we include little more than an abstract and an outline, (ii) work that will not be reported elsewhere, for which much more detail is provided, (iii) work that is not yet ready for publication, for which unpolished, incomplete results are given.

TABLE OF CONTENTS

I. COMPUTATIONAL ERROR ESTIMATES AND DEFERRED CORRECTIONS FOR OPERATOR EQUATIONS

Digital computer simulations of physical phenomena often involve the numerical solution of partial differential equations with initial and boundary conditions. It is clearly important to have an estimate of the accuracy of the computed solutions. We have identified ten ways to estimate the error in the numerical solution, often called the global error. Good theoretical and practical results for estimating error for smooth problems were obtained by Lindberg (1976) using a generalization of Fox's deferred difference correction. The idea is to produce another numerical solution of enhanced accuracy by subtracting out from the difference equation of the improved solution local error estimates computed for the original solution. We have since been able to simplify and strengthen Lindberg's theorems. The result is a theoretical framework for proving accuracy results for deferred correction. Application of this theory to ordinary differential equations yields useful theoretical results. The practical benefit is to offer guidance in the construction of local error estimators by identifying those properties that are important for accurate error estimation. These ideas have been applied to time-dependent partial differential equations with nonsmooth solutions, particularly hyperbolic problems with shock discontinuities. For these problems it would be very difficult to obtain error estimates that can be theoretically justified; we were content to devise and test schemes having only limited theoretical support.

An introduction to the idea of deferred correction and our style of error analysis is given in Skeel (1980, section 1).

1. Error Estimation and Iterative Improvement for the Numerical Solution of Operator Equations (Lindberg; Van Rosendale, Skeel)

A technical report with this title was written by Lindberg (1976). The abstract follows:

A method for estimation of the global discretization error of solutions of operator equations is presented. Further an algorithm for iterative improvement of the approximate solution of such problems is given. The theoretical foundation for the algorithms are given as a number of theorems. Several classes of operator equations are examined and numerical results for both the error estimation algorithm and the algorithm for iterative improvement are given for some classes of ordinary and partial differential equations and integral equations.

The table of contents is as follows:

An improved and shorter version of this report (Lindberg (1980)) is to appear in BIT.

2. A Theoretical Framework for Proving Accuracy Results for Deferred Corrections (Skeel; Ortman)

A manuscript with this title has been submitted for publication by Skeel (1980). The abstract follows:

> General techniques are described for proving accuracy results for deferred correction solutions to differential equations. These techniques apply also to computational estimates of the local discretization error. The proofs avoid the necessity of demonstrating the existence of asymptotic expansions of the global error in powers of some meshsize parameter.

The outline is as follows:

1. Introduction
2. Historical Survey
   - 2.1. Difference Correction
   - 2.2. Difference Estimates of the Local Error
   - 2.3. Generalized Difference Correction
   - 2.4. Defect Estimates of the Local Error
   - 2.5 "Cheap" Global Error Estimates
3. Local Error Properties of Numerical Methods
4. An Error Analysis for One Deferred Correction

The experiment referred to in the penultimate sentence of Section 2.1 of this manuscript was performed in double precision on the CYBER 175 for a decreasing sequence of stepsizes $\Delta t = 1, 1/2, \ldots, 1/128$ for the problem $y' = y$, $y(0) = 1$. The values of the computed solution for $t = 1$ were printed out, and they seemed to be converging very much like $(\Delta t)^4$.

The third paragraph of this manuscript mentions three notable implementations of the deferred correction idea. In addition there is the subroutine SEPELI developed by John Adams at NCAR (see Swarztrauber and Sweet (1979)), which has the option of obtaining fourth-order accurate solutions via deferred corrections applied to a fast separable elliptic PDE solver of second order accuracy. This kind of situation in which we have a very efficient low order method is ideal for the application of deferred corrections.

3. The Order of Accuracy for a Deferred Corrections Algorithm (Skeel)

This work will constitute the future paper mentioned in Skeel (1980, section 4, second paragraph) in which we give an error analysis for a sequence of iterations for the algorithm considered by Christiansen and Russell (1979). This is intended to be a realistic example of the application of the theoretical framework for proving accuracy results, which because of its length was not included in Skeel (1980).

In Christiansen and Russell (1979) a careful analysis of deferred corrections that does not involve asymptotic expansions is done for a realistic algorithm similar to the implementation of Lentini and Pereyra (1977) of iterated deferred corrections for two-point boundary value problems. Under weak assumptions they prove that each iteration increases the order by one, and they give empirical results showing that the first two iterations increase the order by two. They suggest how this might be proved but do not follow through because "Such a proof would be quite tedious." In this paper we sketch a proof of this fact, which we believe is less tedious than that of Christiansen and Russell (1979) due to the way in which we break down the proof into smaller *simply stated* results. The key idea is the use of judiciously chosen "discrete Sobolev" norms for measuring the smoothness of the global error and of the local error. In particular we use special norms like those of Spijker (1971) and Stummel (1975) for the errors. These norms are chosen so that they admit both upper *and lower* bounds on the norms of the global error thus making it feasible to determine the exact order of convergence. However, it must be acknowledged that the algorithm we analyze is a bit simplified and unrealistic for nonlinear problems in that we assume the exact solution of nonlinear equations.

3.1. The numerical method. Assume that the differential operator F given for an arbitrary function z by

$$F(z) := \begin{pmatrix} g(z(0), z(1)) \\ z'(x) - f(x,z(x)),\ 0 \le x \le 1 \end{pmatrix}$$

satisfies the assumptions of Christiansen and Russell (1979) so that in particular the operator equation $F(y) = 0$ has an isolated solution y.

Consider a family of meshes

$$0 = x_0 < x_1 < \ldots < x_J = 1$$

with $h_j := x_j - x_{j-1}$ and $h := 1/J$ such that $h_j \le Ch$ uniformly for all meshes in the given family. (Presumably, the average meshsize h can be arbitrarily close to zero, for otherwise the theorems that follow have no content.) We obtain results for three progressively stronger assumptions on the family of meshes:

(i) "no assumption," meaning that there is no assumption apart from that already stated,

(ii) "weak assumption," meaning that

$$\sum_{j=1}^{J-1} |h_{j+1}/h_j - 1| \le ch$$

uniformly for all meshes, which is quite realistic. Skeel and Jackson (1979) use the term "variation-bounded" to describe such a family of meshes in connection with the stability of multistep methods for initial value problems.

(iii) "strong assumption," meaning that

$$\max_{1 \le j \le J-1} |h_{j+1}/h_j - 1| \le ch$$

uniformly for all meshes, which is not so realistic. The term "locally uniform" is quite descriptive of such a family of meshes.

We seek a numerical solution η_j, $j = 0(1)J$, approximating the theoretical solution on the mesh $y(x_j)$, $j = 0(1)J$. Our discretization will be an approximation of order $2k$ to

$$g(y(x_0), y(x_J)) = 0$$

$$\frac{1}{h_j} \int_{x_{j-1}}^{x_j} y'(x) - f(x,y(x))dx = 0 \ , \ j = 1(1)J \ .$$

With the j-th subinterval $[x_{j-1}, x_j]$ we associate the set of integers $I(j)$ consisting of those $2k$ integers from 0 through J which are nearest $j - 1/2$. We approximate the term $f(x,y(x))$ in the integrand by the polynomial of degree $2k-1$ which interpolates $f(x,y(x))$ on the meshpoints x_i, $i \in I(j)$. Thus centered interpolation is used everywhere except near the boundary. The resulting equation is

$$\frac{1}{h_j} (y(x_j) - y(x_{j-1})) - \sum_{i \in I(j)} \beta_{j,i} \ f(x_i, y(x_i)) = O(h^{2k})$$

where $\beta_{j,i}$ is a function of the relative meshsizes. The numerical solution η_j is the solution of the discrete problem

$$g(\eta_0, \eta_J) = 0 \ ,$$

$$\frac{1}{h_j} (\eta_j - \eta_{j-1}) - \sum_{i \in I(j)} \beta_{j,i} \ f(x_i, \eta_i) = 0 \ , \ j = 1(1)J \ ,$$

which is a system of $J+1$ nonlinear equations whose Jacobian has bandwidth $2k$ times the dimensionality of the original ODE. Thus the cost of solving linear systems involving the Jacobian is proportional to k^2. This can be avoided through iterated deferred corrections.

Let $\zeta = (\zeta_0, \zeta_1, \ldots, \zeta_J)$ be an arbitrary element of the discrete space and define a discrete operator $\phi(\zeta)$ by

$$\phi_k(\zeta)_0 = g(\zeta_0, \zeta_J) \ ,$$

$$\phi_k(\zeta)_j = \frac{1}{h_j} (\zeta_j - \zeta_{j-1}) - \sum_{i \in I(j)} \beta_{j,i} \ f(x_i, \zeta_i) \ , \ j = 1(1)J \ .$$

For k = 1 this is the trapezoid rule

$$\phi_1(\zeta_j) = \frac{1}{h_j}(\zeta_j - \zeta_{j-1}) - \frac{1}{2} f(x_j, \zeta_j) - \frac{1}{2} f(x_{j-1}, \zeta_{j-1}) ,$$

which is the most economical for computation. In iterated deferred corrections, due to Fox (1947), we write

$$\phi_m =: \phi - \psi_m \text{ where } \phi := \phi_1$$

thus expressing ϕ_m as the sum of the economical trapezoid rule plus the correction $-\psi_m$. Instead of solving $\phi(\eta) - \psi_m(\eta) = 0$, the procedure is to solve

$$\phi(\eta^1) = 0$$

and

$$\phi(\eta^k) - \psi_m(\eta^{k-1}) = 0 , \quad k = 2, 3, \ldots .$$

It can be shown that the high order corrections are unnecessary for small k and it is less work to do *iterative updating deferred correction*:

$$\phi(\eta^1) = 0 ,$$

$$\phi(\eta^k) - \psi_k(\eta^{k-1}) = 0 , \quad k = 2(1)m .$$

3.2. Stability. We are interested in the norm of the error $\|\eta^k - \Delta y\|_0$ where $\Delta y = (y(x_0), y(x_1), \ldots, y(x_J))$ and the norm is defined by $\|\zeta\|_0 = \max_{0 \le j \le J} |\zeta_j|$. We can reduce this to the local level because ϕ is *stable*:

$$\|\tilde{\zeta} - \zeta\|_0 \le S_0 \|\phi(\tilde{\zeta}) - \phi(\zeta)\|_{*0}$$

for some S_0 independent of the mesh. And so

$$\|\eta^k - \Delta y\|_0 \le S_0 \|\phi(\eta^k) - \phi(\Delta y)\|_{*0} .$$

The norm for the discrete residual space E_h^{*0} has not yet been specified but the best choice is the Spijker norm

$$\|\rho\|_{*0} := \max_{0\le j\le J} \left|\rho_0 + \sum_{i=1}^{j} h_i\rho_i\right|$$

where ρ is an arbitrary element of the discrete residual space. (This norm also yields a lower bound in the stability definition.) Stability depends crucially on the solution being an isolated solution (Keller (1976)).

If the analysis of the error is pursued, it turns out that accuracy improvement depends on the smoothness of the error. For this purpose we define the following norms for the discrete solution space:

$$\|\zeta\|_1 := \max\{|\zeta_0|, \max_{1\le j\le J} |\dot\zeta_j|\},$$

$$\|\zeta\|_2 := \max\{|\zeta_0|, |\dot\zeta_1|, \max_{2\le j\le J} |\dot\zeta_j|\}$$

where

$$\dot\zeta_j := (\zeta_j - \zeta_{j-1})/(x_j - x_{j-1})$$

and

$$\dot\zeta_j := (\dot\zeta_j - \dot\zeta_{j-1})/(x_j - x_{j-2}) .$$

Define the following norms for the discrete residual space:

$$\|\rho\|_{*1} := \max\{|\rho_0|, \max_{1\le j\le J} |\rho_j|\},$$

$$\|\rho\|_{*2} := \max\{|\rho_0|, |\rho_1|, \max_{2\le j\le J} |\dot\rho_j|\}.$$

With considerable effort one can establish the following stability results:

(S_m) $$\|\tilde\zeta - \zeta\|_m \le S_m \|\phi(\tilde\zeta) - \phi(\zeta)\|_{*m}, \quad m = 0, 1, 2 .$$

The case m = 1 is implied by Christiansen and Russell (1979, lemma 1) for linear problems.

3.3. Local errors. Returning to the norm of the error, we have using the more general norm

$$\|\eta^k - \Delta y\|_m \le S_m \|\phi(\eta^k) - \phi(\Delta y)\|_{*m}$$

$$= S_m \|\psi_k(\eta^{k-1}) - \phi(\Delta y)\|_{*m} .$$

Thus the accuracy of η^k depends on the accuracy of $\psi_k(\eta^{k-1})$ as an estimate of the local error $\phi(\Delta y)$. To underscore this point, we write the algorithm as

$$\phi(\eta^1) = 0 ,$$

$$\phi(\eta^k) = \psi_k(\eta^{k-1}), \quad k = 2, 3, \ldots ,$$

the idea being that if the right hand side were exactly the local error then η^k would be exactly Δy. The error in the local error estimate can be split into the propagated error arising from the error $\eta^{k-1} - \Delta y$ and the local error arising from ϕ_k:

$$\psi_k(\eta^{k-1}) - \phi(\Delta y) = \psi_k(\eta^{k-1}) - \psi_k(\Delta y) + \psi_k(\Delta y) - \phi(\Delta y)$$

$$= [\psi_k(\eta^{k-1}) - \psi_k(\Delta y)] - \phi_k(\Delta y) .$$

Each of these terms is analyzed separately; in this subsection we examine $-\phi_k(\Delta y)$, which is the local error of the formula of order $2k$.

With "no assumptions" we have both

$$(c_0^k) \qquad \|\phi_k(\Delta y)\|_{*0} \le h^{2k} c_0^k , \quad k = 1, 2, \ldots$$

and

$$(c_1^k) \qquad \|\phi_k(\Delta y)\|_{*1} \le h^{2k} c_1^k , \quad k = 1, 2, \ldots .$$

However this is not true for the E_h^{*2} norm for $k \ge 2$ even for uniform mesh families because the use of uncentered formulas at the two ends causes abrupt changes in the local error there. With the "strong assumption" one can show that

(c_2) $$\|\phi(\Delta y)\|_{*2} \le h^2 c_2 .$$

Under weaker assumptions this is not true because if the meshsize does not vary slowly then neither does the local error.

3.4. Contractivity. The other term in our error analysis is $\|\psi_k(\eta^{k-1}) - \psi_k(\Delta y)\|_{*m}$. If there is to be an increase in the order of accuracy, the operator ψ_k needs to be contractive with contractive power at least O(h). As an example,

$$\psi_2(\zeta)_j = \frac{h_j^2}{12}\left(\frac{h_j+2h_{j+1}}{h_{j-1}+h_j+h_{j+1}}\,\psi^2 f(x_j,\zeta_j) + \frac{h_j+2h_{j-1}}{h_{j-1}+h_j+h_{j+1}}\,\psi^2 f(x_{j+1},\zeta_{j+1})\right) .$$

We note that at best

$$|\psi_2(\tilde{\zeta})_j - \psi_2(\zeta)_j| = O(\|\tilde{\zeta} - \zeta\|_0)$$

so that contractivity is only O(1). However, if we use the norm for the discrete solution space E_h^2 involving second order divided differences, then

$$|\psi_2(\tilde{\zeta})_j - \psi_2(\zeta)_j| = O(h^2 \|\tilde{\zeta} - \zeta\|_2) .$$

This second "bound" is better than the first *if* ζ is smooth. We shall prove contractivity results of two types:

$$\|\psi_k(\tilde{\zeta}) - \psi_k(\zeta)\|_{*m} \le hL_{mm}^k \|\tilde{\zeta} - \zeta\|_m$$

and

$$\|\psi_k(\tilde{\zeta}) - \psi_k(\zeta)\|_{*m} \le h^2 L_{m,m+1}^k \|\tilde{\zeta} - \zeta\|_{m+1} .$$

Note that more contractivity is possible if we are willing to go to a stronger norm.

The polynomial interpolant used to derive ϕ_k and hence ψ_k can be expressed in Newtonian form in terms of divided differences. One would find that for $1 \le j \le J$

$$\psi_k(\zeta)_j = \sum_{i=2}^{2k-1} \gamma_{j,i} \, \nabla^i f(x_{p(j)}, \zeta_{p(j)})$$

where

$$p(j) = \begin{cases} 2k-1 & , \ 1 \le j \le k \ , \\ j+k-1 & , \ k \le j \le J-k+1 \ , \\ J & , \ J-k+1 \le j \le J \ , \end{cases}$$

and the $\gamma_{j,i}$ depend only on the relative meshsizes.

THEOREM 1. *The following contractivity results hold for the* E_h^{*0} *norm on uniform meshes:*

$$(L_{00}^k) \qquad \|\psi_k(\tilde{\zeta}) - \psi_k(\zeta)\|_{*0} \le h L_{00}^k \, \|\tilde{\zeta} - \zeta\|_0 \ ,$$

$$(L_{01}^k) \qquad \|\psi_k(\tilde{\zeta}) - \psi_k(\zeta)\|_{*0} \le h^2 L_{01}^k \, \|\tilde{\zeta} - \zeta\|_1 \ .$$

Proof. We have that

$$\psi_k(\zeta)_i = h^2 \sum_{\ell=2}^{J} \alpha_{i,\ell} \, \nabla^2 f(x_\ell, \zeta_\ell)$$

where $\alpha_{i,\ell} = 0$ for $\ell < \min \{i-k+2, J-2k+3\}$ and $\ell > \max \{i+k-1, 2k-1\}$. It is sufficient to show that

$$h\psi_k(\zeta)_j \ , \ j = 1(1)k-1, \ J-k+2(1)J$$

and

$$h \sum_{i=k}^{j} \psi_k(\zeta)_i \ , \ j = k(1)J-k+1$$

satisfy a Lipschitz condition of the form given in the statement of the theorem. We have

$$h\psi_k(\zeta)_j = h^2 \sum_{\ell=2}^{2k-1} \alpha_{j,\ell} \, [\nabla f(x_\ell, \zeta_\ell) - \nabla f(x_{\ell-1}, \zeta_{\ell-1})]$$

and

$$h \sum_{i=k}^{j} \psi_k(\zeta)_i = h^2 \sum_{i=k}^{j} \sum_{\ell=i-k+2}^{i+k-1} \alpha_{i,\ell} \nabla f(x_\ell, \zeta_\ell)$$

$$- h^2 \sum_{i=k-1}^{j-1} \sum_{\ell=i-k+2}^{i+k-1} \alpha_{i+1,\ell+1} \nabla f(x_\ell, \zeta_\ell)$$

$$= h^2 \sum_{\ell=j-k-2}^{j+k-1} \alpha_{j,\ell} \nabla f(x_\ell, \zeta_\ell) - h^2 \sum_{\ell=1}^{2k-2} \alpha_{k,\ell+1} \nabla f(x_\ell, \zeta_\ell) .$$

Now it is enough to show that $h^2 \nabla f(x_j, \zeta_j)$ satisfies a Lipschitz condition of the given form. For the first Lipschitz condition this follows readily from the Lipschitz continuity of f. For the second inequality note that

$$\nabla f(x_j, \zeta_j + \xi_j) - \nabla f(x_j, \zeta_j)$$

$$= \nabla \{ \int_0^1 f_y(x_j, \zeta_j + \theta\xi_j) d\theta \cdot \xi_j \}$$

$$= \int_0^1 \nabla f_y(x_j, \zeta_j + \theta\xi_j) d\theta \cdot \xi_j + \int_0^1 f_y(x_{j-1}, \zeta_{j-1} + \theta\xi_{j-1}) d\theta \cdot \nabla \xi_j \quad . \square$$

Remark. This theorem actually holds under the "weak assumption." It may be possible to further weaken the assumption on meshes for the first inequality by noting that only $\alpha_{i\ell} - 2\alpha_{i+1,\ell+1} + \alpha_{i+2,\ell+2}$ need be small rather than

$$\alpha_{i\ell} - \alpha_{i+1,\ell+1} .$$

THEOREM 2. *The following contractivity results hold for the* E_h^{*1} *norm with "no assumption":*

$$(L_{11}^k) \qquad \|\psi_k(\tilde{\zeta}) - \psi_k(\zeta)\|_{*1} \le h L_{11}^k \|\tilde{\zeta} - \zeta\|_1 ,$$

$$(L_{12}^k) \qquad \|\psi_k(\tilde{\zeta}) - \psi_k(\zeta)\|_{*1} \le h^2 L_{12}^k \|\tilde{\zeta} - \zeta\|_2 .$$

Proof. It is sufficient to show that $\psi_k(\zeta)_j$ satisfies the given Lipschitz conditions and this we do as in the proof of Theorem 1. $\square$

Remark. It may be possible to combine a contractivity result for ψ_k with a stability result for ϕ in order to establish a stability result for ϕ_k.

3.5. Order of convergence. We separately consider the results for "strong assumption," "weak assumption," and "no assumption."

With the "strong assumption" it immediately follows from the results of sections 2, 3, and 4 that

$$\|\eta^1 - \Delta y\|_2 \le S_2 h^2 c_2 ,$$

$$\|\eta^2 - \Delta y\|_1 \le S_1 (h^2 L_{12}^2 \|\eta^1 - \Delta y\|_2 + h^4 c_1^2) ,$$

$$\|\eta^3 - \Delta y\|_0 \le S_0 (h^2 L_{01}^3 \|\eta^2 - \Delta y\|_1 + h^6 c_0^3) ,$$

and for k = 4, 5, 6,...

$$\|\eta^k - \Delta y\|_0 \le S_0 (h L_{00}^k \|\eta^{k-1} - \Delta y\|_0 + h^{2k} c_0^k)$$

so that the order progression is 2, 4, 6, 7, 8, 9,.... The result actually proved for this algorithm by Christiansen and Russell is that the order progression is 2, 4, 5, 6, 7, 8,... although they give empirical evidence for the stronger result and outline the proof for this result under the stronger assumption that $h_{j+1}/h_j = 1 + O(h^2)$.

With the "weak assumption" we have

$$\|\eta^1 - \Delta y\|_1 \le S_1 h^2 c_1^1 ,$$

$$\|\eta^2 - \Delta y\|_0 \le S_0 (h^2 L_{01}^2 \|\eta^1 - \Delta y\|_1 + h^4 c_0^2) ,$$

and for k = 3, 4, 5, 6,...

$$\|\eta^k - \Delta y\|_0 \le S_0 (h L_{00}^k \|\eta^{k-1} - \Delta y\|_0 + h^{2k} c_0^k)$$

so that the order progression is 2, 4, 5, 6, 7, 8....

With "no assumption" we have

$$\|\eta^1 - \Delta y\|_1 \le S_1 h^2 c_1^1 ,$$

and for k = 2, 3, 4, 5, 6,...

$$\|\eta^k - \Delta y\|_1 \le S_1 (h L_{11}^k \|\eta^{k-1} - \Delta y\|_1 + h^{2k} c_1^k)$$

so that the order progression is 2, 3, 4, 5, 6, 7,....

4. Ten Ways to Estimate Global Error (Skeel)

A manuscript with this title is being prepared for publication.

An outline follows:

0. Introduction. We describe and assess global error estimation techniques.

1. Deferred correction.

2. Linearized deferred correction.

3. Differential correction.

4. Linearized differential correction.

5. Defect correction. Conditions are established sufficient for the validity of the defect correction idea of P.E. Zadunaisky, and it is shown that this approach offers no theoretical advantage over deferred correction.

6. Richardson extrapolation.

7. Error-gradient estimation. This recent idea is due to Epstein and Hicks (1979).

8. Using two different tolerances.

9. Using two different methods.

10. Using a method with a specially chosen form of error. This idea of Stetter (1971) has a weakness which has been described to us by F.T. Krogh.

5. Nonsmooth Situations (Skeel)

We consider the use of a single deferred correction to obtain an improved solution and hence an estimate of the global error. Recall that there are three components to such a computation:

(i) a numerical solution η which approximates the restriction to a mesh Δy of the theoretical solution to the operator equation $F(y) = 0$.

(ii) a discretization ϕ of F which is computationally attractive.

(iii) a local error estimator ψ for ϕ.

A corrected solution $\bar{\eta}$ is obtained from $\phi(\bar{\eta}) = \psi(\eta)$. The theoretical justification for this procedure is valid only under certain smoothness assumptions on the original solution η and on the problem F. Numerical experiments have demonstrated the power of this technique when the assumptions are satisfied, but when they are not, the quality of the estimates can deteriorate, in particular, the order of accuracy predicted by the theory will not materialize. This section discusses in general terms what might be done in nonsmooth situations. Application to parabolic PDEs and to hyperbolic PDEs is the subject of sections 6 and 7.

Of special interest are hyperbolic PDEs with shocks, and for these only very low orders of accuracy seem possible by any method with the possible exception of very complicated shock fitting methods. This indicates that it would be difficult to construct asymptotically correct estimates for the local error near discontinuities in the solution. Hence accuracy considerations for error estimation techniques must go beyond the order of accuracy. For good estimates in smooth regions one would probably desire an asymptotically correct estimate as $h \to 0$. Among all such estimators we would like to select an (inexpensive) estimator, which is still somewhat accurate when the solution is nonsmooth.

It is not entirely clear how to measure the error so that $\|\eta - \Delta y\|_0$ is small for accurate solutions, since a slightly displaced shock gives a large local value of the global error $\varepsilon := \eta - \Delta y$. It may be better to use a root-mean-square or mean norm instead of the max norm for $\|\cdot\|_0$. Also, it might help to define Δy as average values on cells rather than point values, as in the "control volume approach" described by Roache (1975) for deriving finite difference schemes. For second order schemes the order is the same regardless of whether we are trying to compute average values or point values, but it does matter for higher order schemes.

A careful examination of the theoretical basis of the error estimation technique is done by Skeel (1980) with the aim of identifying the key assumptions necessary for the success of this technique. In particular, assumptions involving asymptotic expansions in the gridsize h were avoided. It was found that the discrepancy $\bar{\eta} - \Delta y$ in the global error estimate satisfies the error bound

$$\|\bar{\eta} - \Delta y\|_0 \leq S(h^p K \|\varepsilon\|_q + \|\psi(\Delta y) - \phi(\Delta y)\|_{*0})$$

where S is the stability constant in

$$\|\tilde{\zeta} - \zeta\|_0 \leq S \|\phi(\tilde{\zeta}) - \phi(\zeta)\|_{*0}$$

and Kh^p is the contractivity constant in

$$(*) \qquad \|\psi(\tilde{\zeta}) - \psi(\zeta)\|_{*0} \leq h^p K \|\tilde{\zeta} - \zeta\|_q .$$

Thus there are four factors that affect the accuracy of the global error estimate. The first factor is the norm of the error $\|\varepsilon\|_q$, which is defined in terms of the first q divided differences of ε. It would be desirable to modify the original numerical solution η before estimating the local error so as to reduce the norm of the error. This may be

possible if one knows something about the behavior of ε, for example that ε alternates in sign. The second factor is the contractivity $h^p K$ of the local error estimator ψ. Very often ψ is chosen to be $\phi - \bar{\phi}$ where $\bar{\phi}$ is a more accurate discretization of F. In such cases one should choose $\bar{\phi}$ to be close to ϕ in some sense. The third factor is the accuracy of the local error estimator $\psi(\Delta y) - \phi(\Delta y)$. The fourth factor is the stability constant S of the method ϕ. In the remainder of this section, we discuss each of these in more detail.

Accuracy of the original solution. One of the contributions to the inaccuracy of the local error estimate comes from the original global error and its first q divided differences. The local error estimate will not be accurate if the error is not a smooth "function" of the independent variables. There are potentially numerous reasons for roughness in the global error:

(i) roundoff error. This would normally be insignificant for those problems of interest.

(ii) iteration error. This arises from the use of implicit difference schemes and can be minimized by doing enough iterations.

(iii) the use of more than one difference scheme to approximate the differential equation. This situation occurs when two-level difference schemes are used to calculate starting values for three-level schemes; it also occurs in hyperbolic equations when implicit schemes are used to calculate numerical boundary values for explicit schemes.

(iv) coarseness of the grid. Certain problems such as stiff differential systems can be accurately solved even though the grid is too coarse for the asymptotic theory to hold.

(v) irregularly spaced gridpoints.

(vi) nonsmoothness in the problem iteslf including the boundary.

If something is known about the behavior of the error, it is possible by smoothing the numerical solution η to reduce the error and especially the divided differences of the error. For example, in the case of shock calculations, higher order schemes introduce systematic oscillatory errors and random choice methods introduce random errors into the solution. In our operator notation this means the decomposition $\psi = \hat{\psi} \cdot \chi$ into a smoothing operator χ and a simpler local error estimator $\hat{\psi}$. Experimentation on a simple test problem with 1/4, 1/2, 1/4 smoothing in space resulted in error reduction near the shock but not away from the shock. It should be possible to refine such techniques by the use of a switch triggered, for example, by sign changes in second differences of the numerical solution. A related approach is used by Lindberg (1976) for the leap-frog method. There the local error estimate for a gridpoint was constructed from computed values only at every other gridpoint. Either of these approaches is applicable to stiff ordinary differential equations solved by the trapezoidal rule. If less is known about the error behavior, one can treat the roughness as noise and construct approximations, such as least-squares approximations, which filter out the noise.

Contractivity of the local error estimator. One basis for evaluating potential local error estimators is inequality (*). One determines the weakest differentiability assumptions and the smallest numerical constant $h^p K$. This can also be done for reduced integer values of p and/or q. Under weaker smoothness assumptions it is quite acceptable that the contractivity of ψ be some fraction of unity rather than O(h).

It should also be fruitful to study the contractivity for simple test problems, in particular, the constant coefficient linear problem. This can be used to determine the contractivity of ψ or better still the contractivity of $\psi \cdot \phi^{-1}$, since the Frechet derivative of the later operator maps the original local errors into inaccuracies in the local error estimate. Following is an example of this type of analysis in a situation where asymptotic analysis in powers of h is inappropriate:

Example. Consider a stiff system of ODEs $y' = f(t, y)$ with initial conditions. Let $\hat{\phi}$ be the finite difference operator for the backward Euler method. Define the global errors $\varepsilon_n := \eta_n - y(t_n)$ and the local errors

$$\delta_0 := 0, \ \delta_n := -(y(t_n) - y(t_{n-1}))/h + f(t_n, y(t_n)) \ .$$

Consider the local error estimators

$$\psi^A(\eta)_n := \frac{h}{2} \left(f(t_n, \eta_n) - f(t_{n-1}, \eta_{n-1})\right)/h$$

and

$$\psi^B(\eta)_n := \frac{h}{2} f^{(1)}(t_n, \eta_n) \ .$$

Then it can be shown that for the test problem $y' = \lambda y + g(t)$, $\mathrm{Re}\ \lambda \leq 0$,

$$\psi^A(\Delta y + \varepsilon)_n - \psi^A(\Delta y)_n = \frac{h\lambda}{2} \sum_{j=1}^{n} (1 - h\lambda)^{n-j+1} (\delta_j - \delta_{j-1}) =: \rho_n \ ,$$

which is small compared to δ_n, since

$$|\rho_n| \leq \frac{1}{2} n^{-1/2} |\delta_1| + \frac{h}{2} \left(\frac{27}{16}\right)^{1/4} \max_{2 \leq j \leq n} |(\delta_j - \delta_{j-1})/h| \ .$$

However,

$$\psi^B(\Delta y + \varepsilon)_n - \psi^B(\Delta y)_n = \rho_n - \frac{h\lambda}{2} \delta_n \ ,$$

which can be very large compared to δ_n.

Accuracy of the local error estimator. A potential estimator ψ can also be evaluated by determining the weakest differentiability assumptions and the smallest numerical constant for a bound on $\|\psi(\Delta y) - \phi(\Delta y)\|_{*0}$. This can also be done for lower orders of accuracy.

Stability of the method. For global error estimation it may be worthwhile to consider methods ϕ for the "improved" solution which are different from the original method $\hat{\phi}$. For example, for conservation laws one might choose a good monotone method for ϕ since it may respond in a more stable fashion to the perturbations $\psi(\eta)$ that are introduced into the second integration.

6. Parabolic Problems (Skeel; Van Rosendale)

A couple of nonsmooth problems of this type were tested by Lindberg (1976): the heat equation

$$u_t = u_{xx}$$

with nonsmooth initial conditions and Burgers' equation

$$u_t + uu_x = \nu u_{xx}$$

with a small value for ν. This second problem develops a sharp front with large spatial derivatives. The local error estimators of Lindberg have the form $\psi^E := \phi - \phi^E$ where ϕ^E is a higher order discretization of the operator based on polynomial interpolation. We attempted to improve his results by using a higher order discretization $\bar{\phi}$ having a more compact computational molecule more like that of ϕ. This would hopefully yield a more contractive $\psi := \phi - \bar{\phi}$.

Statistics were generated for both local and global error estimates at each meshpoint in space-time. These included not only the correct error and the inaccuracy of the error estimate but also the two separate contributions to the inaccuracies: the part inherited from the error of the original numerical solution and the part introduced by the local error estimator. These two contributions were evaluated by splitting the inaccuracy of the local error estimate $\psi(\eta) - \phi(\Delta y)$ into $\psi(\eta) - \psi(\Delta y)$ plus $-\bar{\phi}(\Delta y)$. Perhaps it would have been better to split the inaccuracy of the *global* error estimate according to

$$\bar{\eta} - \Delta y = (\bar{\eta} - \ddot{\eta}) + (\ddot{\eta} - \Delta y)$$

where $\ddot{\eta}$ solves

$$\phi(\ddot{\eta}) = \psi(\Delta y) \ .$$

Thus we could have separately evaluated the centractivity of ψ given by

$\|\bar{\eta} - \ddot{\eta}\|/\|\eta - \Delta y\|$ and the relative error of ψ, which is in some sense measured by $\|\ddot{\eta} - \Delta y\|/\|\eta - \Delta y\|$. This information might be meaningful even if ψ is only to be used for local error estimation.

Results for our compact local error estimators are better than for Lindberg's estimators in regions of smoothness but are as bad at discontinuities.

6.1. Heat equation. The differential equation

$$u_t - u_{xx} = 0$$

was solved with zero boundary conditions and with both smooth initial conditions

$$u(x, 0) = \sin x \quad , \quad 0 \leq x \leq \pi \ ,$$

and nonsmooth initial conditions

$$u(x, 0) = \min \{x, \pi - x\} \quad , \quad 0 \leq x \leq \pi \ .$$

Analytic solutions are given by Lindberg (1976).

The original numerical solution U was obtained by FTCS differencing

$$\frac{1}{k} \delta_t U_j^{n+1/2} - \frac{1}{h^2} \delta_x^2 U_j^n = 0$$

where $U_j^n \approx u(jh, nk)$. This is second order in the spatial gridsize $h = 1/J$ provided that the stability condition $k \leq h^2/2$ is satisfied. This scheme was also used as the basic method ϕ for obtaining the corrected solution.

Lindberg (1976) considers $\psi^E := \phi - \phi^E$ where ϕ^E is based on quadratic interpolation in time and quartic interpolation in space

$$\frac{1}{k} \mu_t \delta_t U_j^n - \frac{1}{h^2} (\delta_x^2 - \frac{1}{12} \delta_x^4) U_j^n = 0$$

with modifications for $j = 1$, $j = J-1$ and $n = 1$. We also consider the

operator $\bar{\phi}$ given by

$$\frac{1}{k}\left(1 + \frac{1}{12}\delta_x^2\right)\delta_t U_j^{n+1/2} - \frac{1}{h^2}\delta_x^2 \mu_t U_j^{n+1/2} = 0 ,$$

because it is the method fourth order in h which differs the least from ϕ. The local error estimator $\psi^E := \phi - \phi^E$ is given for arbitrary V by

$$\frac{k}{2}\frac{1}{k^2}\delta_t^2 V_j^n - \frac{h^2}{12}\frac{1}{h^4}\delta_x^4 V_j^n$$

and $\psi := \phi - \bar{\phi}$ is given by

$$\left(\frac{k}{2} - \frac{h^2}{12}\right)\frac{1}{kh^2}\delta_t\delta_x^2 V_j^{n+1/2} .$$

We are interested in how contractive these operators are. Since they are both linear, it is sufficient to consider $\|\psi^E(V)\|$ and $\|\psi(V)\|$ where the norm is the max norm. We obtain the bounds

$$\|\psi^E(V)\| \le \tfrac{1}{3}|V|_{xx} + |V|_t$$

and

$$\|\psi(V)\| \le \left|\frac{h^2 - 6k}{3h^2k}\right| \min\left\{\frac{h^2}{2}|V|_{xx},\ k|V|_t\right\}$$

where $|V|_{xx}$ is the maximum of all $|h^{-2}\delta_x^2 V_j^n|$ and $|V|_t$ is the maximum of all $|k^{-1}\delta_t V_j^n|$. One can show that

$$(1) \qquad \frac{\text{bound for } \psi}{\text{bound for } \psi^E} \le \left|\frac{h^2 - 6k}{3h^2 + 2k}\right| \le \frac{1}{2}$$

assuming that $k \le h^2/2$. With even weaker norms we get

$$\|\psi^E(V)\| \le \frac{6h^2 + 4k}{3h^2k}\|V\|$$

and

$$\|\psi(V)\| \le \left|\frac{2h^2 - 12k}{3h^2k}\right|\|V\| ,$$

and again (1) holds for this last pair of bounds. This suggests that $\bar{\phi}$ is more contractive than ϕ^E. It is worth noting that $\bar{\phi}$ becomes identical to ϕ if $k = h^2/6$ with the result that the error estimates would be zero! Nonetheless, zero would be a good approximation to the error *in an absolute error sense.*

The table that follows compares for the smooth problem the error estimates at x_4 for a uniform grid $0 = x_0 < x_1 < \ldots < x_9 = \pi$ with $k = (5/2\pi^2)h^2$. The estimated global errors are $U - U^E$ and $U - \bar{U}$ where U^E solves $\phi(U^E) = \psi^E(U)$ and $\bar{U}$ solves $\phi(\bar{U}) = \psi(U)$. The errors in the local error estimates $\psi^E(U)$ and $\psi(U)$ can be split into two parts:

$$\psi^E(U) - \phi(\Delta u) = [\psi^E(\Delta u + \varepsilon) - \psi^E(\Delta u)] - \phi^E(\Delta u) ,$$

$$\psi(U) - \phi(\Delta u) = [\psi(\Delta u + \varepsilon) - \psi(\Delta u)] - \bar{\phi}(\Delta u) .$$

The first part is due to the error $\varepsilon := U - \Delta u$ in the original solution U, and the second part is the local error introduced by the more accurate operators $\phi^E = \phi - \psi^E$ and $\bar{\phi} = \phi - \psi$.

time level	global error	error in global error estimate		local error	error in local error estimate = part due to ε + new error	
		for ψ^E	for ψ		for ψ^E	for ψ
1	-157	0	0	-5083	-10 135 -145	-10 7 -17
2	-304	-15	-1	-4929	-464 -771 307	-9 7 -16
4	-572	-40	-1	-4634	-434 -723 289	-7 8 -15
8	-1011	-77	-1	-4096	-381 -637 255	-3 10 -13

Note: all values are 10^6 times actual values.

The next table displays the same information for the nonsmooth problem:

time level	global error	error in global error estimate		local error	error in local error estimate = part due to ε + new error	
		for ψ^E	for ψ		for ψ^E	for ψ
1	-177	139	139	-5741	4500 12355 -7855	4500 9294 -4793
2	-147	54	110	-745	-1705 251 -1956	133 161 -28
4	-114	51	81	-308	-295 -595 300	41 48 -7
8	-84	47	58	-116	-52 -102 50	10 11 -2

Note: all values are 10^4 times actual values.

Hence for both problems ψ was a much more accurate local error estimator than ψ^E although the big local error near the point of discontinuity was still underestimated by a factor of 4.6. The better global error estimator was ψ^E due to a fortuitous cancellation of errors in the local error estimates resulting from the use of a special local error estimate for the first time level. Apparently the two parts of the error in the local error estimate tend to cancel especially in the case of the smooth problem. It is not clear whether or not this desirable situation can be caused to happen more generally. The foregoing experiments were performed also for $h = \pi/19$ and $h = \pi/39$. The qualitative nature of the results seemed to depend on the time level n rather than the time t.

The previously tabulated calculations were also performed for the uniform grid $0 = x_0 < x_1 < \ldots < x_{10} = \pi$ with $k = (5/2\pi^2)h^2$. Error estimates are given at x_5, which lies right on the discontinuity:

time level	global error	error in global error estimate for ψ^E	error in global error estimate for ψ	local error	error in local error estimate = part due to ε + new error for ψ^E	error in local error estimate = part due to ε + new error for ψ
1	193	-330	-330	7703	-13218 -31213 17995	-13218 -26542 13324
2	146	-519	-213	-1235	-12344 -29350 17006	-90 -129 40
4	106	-203	-148	-420	-735 -1262 527	-1 8 -9
8	76	-120	-104	-142	-95 -161 66	0 3 -2

Note: all values are 10^4 times actual values.

We note that the global error estimates are really bad due to the very poor local error estimate at the initial discontinuity. Otherwise ψ was much more accurate than ψ^E.

Clearly a nonuniform mesh would be appropriate for the nonsmooth problem, and so the grid points

$$x_j = \frac{\pi}{2}\left[\left(4\left(\frac{j}{J} - \frac{1}{2}\right)^2 + 1\right)\left(\frac{j}{J} - \frac{1}{2}\right) + 1\right]$$

were used with J = 10 and the time increment

$$k = \frac{1}{5}\left(\min_j x_j - x_{j-1}\right)^2 .$$

The gridpoint density is four times as great at the center as at the ends. The method ϕ is given by

$$\frac{\delta}{\delta t} U_j^{n+1/2} - \frac{\delta^2}{\delta x^2} U_j^n = 0$$

where

$$\frac{\delta}{\delta t} U_j^{n+1/2} = \frac{1}{k}\left(U_j^{n+1} - U_j^n\right)$$

and

$$\frac{\delta}{\delta x} U_j^n = \frac{1}{2} \left(\frac{1}{h_{j+1}} (U_{j+1}^n - U_j^n) + \frac{1}{h_j} (U_j^n - U_{j-1}^n) \right) ,$$

which is generally only first order in space. The higher order method $\bar{\phi}$ is given by

$$\left(1 + \frac{1}{12} h_j h_{j+1} \frac{\delta^2}{\delta x^2} + \frac{1}{3} (h_{j+1} - h_j) \frac{\delta}{\delta x} \right) \frac{\delta}{\delta t} U_j^{n+1/2} - \frac{\delta^2}{\delta x^2} U_j^{n+1/2} = 0$$

where

$$U_j^{n+1/2} = \frac{1}{2} (U_j^n + U_j^{n+1})$$

and

$$\frac{\delta}{\delta x} U_j^n = \frac{1}{2} \left(\frac{1}{h_{j+1}} (U_{j+1}^n - U_j^n) + \frac{1}{h_j} (U_j^n - U_{j-1}^n) \right) ,$$

which is generally only third order in space. Hence the local error estimator $\psi := \phi - \bar{\phi}$ is given by

$$\left(\frac{k}{2} - \frac{1}{12} h_j h_{j+1}\right) \frac{\delta}{\delta t} \frac{\delta^2}{\delta x^2} V_j^{n+1/2} - \frac{1}{3} (h_{j+1} - h_j) \frac{\delta}{\delta t} \frac{\delta}{\delta x} V_j^{n+1/2} .$$

Numerical results are given below for $x_5 = \pi/2$.

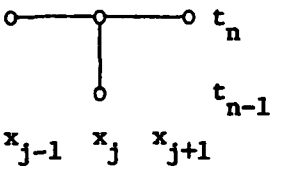
time level	global error	error in global error estimate for ψ	local error
2	120	-133	-434
8	82	-59	-135
20	80	-36	-35
30	83	-29	-19

Note: all values are 10^4 times actual values.

There is an improvement compared to the uniform mesh due to the smaller contribution to the global error from the local error at the discontinuity.

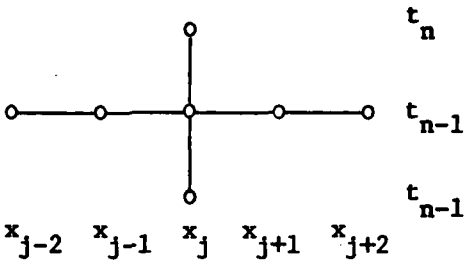
6.2. Burgers' equation. Burgers' equation $u_t + uu_x = \nu u_{xx}$ with initial condition $u(0, x) = \sin x$, $0 < x < \pi$, and zero boundary conditions was solved with an implicit scheme

ϕ :



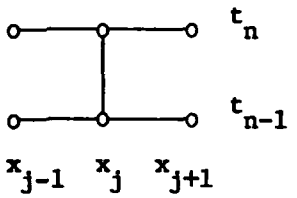
which is first order in k and second order in h. Two difference operators were considered as local error estimators:

ϕ^E :



with modifications for $j = 1$, $j = J-1$, and $n = 1$; and

$\bar{\phi}$:



The difference operator ϕ^E, which is second order in k and fourth order in h, was used by Lindberg; and $\bar{\phi}$, which is second order in both k and h, was chosen because it differed the least from ϕ. The following tables compare the two local error estimates with $h = \pi/20$ and $k = 1/10$ for $\nu = 1$ and for $\nu = 1/16$. In this problem a front develops and then dissipates. The sharpness of the front increases as $\nu \to 0$. From the table it appears that ϕ^E is better than $\bar{\phi}$ except at the first time level where unsymmetric formulas had to be used. The tables give the maximum absolute value of each quantity.

$\nu = 1$

time level	global error	error in global error estimate for ψ^E	error in global error estimate for ψ	local error	error in local error estimate = part due to ε + new error for ψ^E	error in local error estimate = part due to ε + new error for ψ
1	107	35	8	1483	510 704 199	113 134 118
5	188	32	20	399	47 74 29	95 139 43
10	195	20	20	211	25 30 7	32 45 14
20	143	11	4	73	2 4 2	2 1 2
40	39	1	1	10	1 0 0	1 1 0

Note: all values are 10^4 times actual values.

$\nu = 1/16$

time level	global error	error in global error estimate for ψ^E	error in global error estimate for ψ	local error	error in local error estimate = part due to ε + new error for ψ^E	error in local error estimate = part due to ε + new error for ψ
1	45	12	2	478	140 183 42	27 49 41
5	184	45	11	398	112 179 67	68 82 62
10	262	53	205	962	132 462 448	824 1611 806
20	1839	1039	2478	5179	3620 1508 2112	5529 10946 5417
40	1054	340	931	1104	179 401 580	1063 2188 1125
80	213	15	129	72	26 27 2	66 136 70

Note: all values are 10^4 times actual values.

7. Hyperbolic Problems (Skeel; Ortman, Van Rosendale)

Guided by the general theory of section 2 and the ideas of section 5, various numerical schemes were constructed and tested.

The estimation of global discretization error by deferred corrections requires two parallel integrations of the problem, the second integration being corrected by subtracting out local error estimates obtained from the first integration. Thus there are three major components of any numerical test: the initial value problem, the integrator, and the local error estimator.

Recently Stetter (1978) has observed that for global error estimation, the second integration could be performed by the cheapest available (stable and consistent) method. However, there are economies associated with re-using the same method, and so it is not clear in general which approach is more efficient. In all our tests the two integrations are performed with the same numerical method.

Also, we note that Hackbusch (1977) and authors cited therein have applied Richardson extrapolation to methods for hyperbolic systems of first order.

Three ways of constructing the local error estimator ψ are outlined in Skeel (1980, section 4). We were attracted to the Fox-Stetter-Lindberg idea of using $\psi := \phi - \bar{\phi}$ where ϕ is the basic integrator and $\bar{\phi}$ is a more accurate discretization. This permits the use in ψ of specialized techniques for nonlinear hyperbolic problems such as artificial viscosity, upwind differencing, artificial compression, antidiffusion, and hybridization. It is worth noting that as a solution operator, $\bar{\phi}$ need not be stable in order for it to be useful for local error estimation. In addition, the computational cost of "inverting" $\bar{\phi}$ is irrelevant since only $\bar{\phi}(\eta)$ need

be computed. Hence, certain types of implicitness are computationally inexpensive.

Three initial value problems with known solutions were tested: (i) $u_t + u_x = 0$ with a periodic square wave solution, (ii) the inviscid Burgers equation $u_t + (u^2/2)_x = 0$ with initial conditions chosen so that a shock develops, and (iii) the Riemann problem for the system of three equations for one-dimensional Eulerian gas dynamics as in Sod (1977).

The first method ϕ considered was the two-step Lax-Wendroff scheme. A variable-mesh scheme was desired, but it was shown that there is no second order extension of Lax-Wendroff to variable meshes. A reasonably compact two-level fourth order error estimator for Lax-Wendroff was sought but only third order estimators could be found that were also economical. Finally, smoothing schemes for Lax-Wendroff solutions were tested with partial success. The aim was to make the numerical solution more amenable to global error estimation.

Also tested was Lax-Wendroff with artificial viscosity as in Sod (1977).

Since most methods have a fractional order of convergence in the ℓ_1 norm when applied to problems with shocks, it was thought worthwhile to test first order difference schemes. Deferred correction applied to FTCS (forward time centered space) differencing with a Lax-Wendroff local error estimator yielded a solution which was better than that obtained by Lax-Wendroff alone. When Lax's method was teamed up with Lax-Wendroff, the deferred correction solution was better than the original Lax solution but not as good as Lax-Wendroff. The problems tested were $u_t + u_x = 0$ and $u_t + uu_x = 0$.

A number of techniques for hyperbolic systems of conservation laws seem to be motivated by considerations appropriate to a single conservation law. These involve nonsmooth functions of differences such as absolute value, sign, and maximum. Systematic extension of these schemes to systems $\underline{u}_t + \underline{f}(\underline{u})_x = \underline{0}$ is possible: one way is to perform a local transformation which diagonalizes $\underline{f}'(u)$ at some local value of $\underline{u}$ making the system only weakly coupled locally, then to apply the scalar scheme, and finally to reverse the transformation. These transformations can be done implicitly. For example, for upwind differencing this might amount to

$$\frac{1}{k}\,(\underline{U}_j^{n+1} - \underline{U}_j^n) + \frac{1}{2h}\,(\underline{f}(\underline{U}_{j+1}^n) - \underline{f}(\underline{U}_{j-1}^n))$$

$$= \frac{h}{2}\,\mathrm{sgn}(\underline{f}'(\underline{U}_j^n))\,\frac{1}{h^2}\,(\underline{f}(\underline{U}_{j+1}^n) - 2\underline{f}(\underline{U}_j^n) + \underline{f}(\underline{U}_{j-1}^n))$$

where $\mathrm{sgn}(\underline{f}'(\underline{u}))$ is the matrix obtained by diagonalizing $\underline{f}'(u)$, applying the sign function individually to each eigenvalue, and reversing the similarity transformation. This idea is used by Lax and Wendroff (1960) for artificial viscosity, and they show how to do it given the eigenvalues, but not the eigenvectors, of $\underline{f}'(\underline{u})$. Usually the eigenvalues of $\underline{f}'(\underline{u})$ are known, and in any case, there is always the possibility of approximating $\mathrm{sgn}(\underline{f}'(\underline{u}))$ without reducing the order of accuracy. We tested the foregoing upwind scheme on the Eulerian gas dynamics equations, and it performed much better than the upwind differencing tested by Sod (1977). This idea may be useful for artificial compression, antidiffusion, artificial dissipation, and SHASTA.

Deferred corrections can also be used as a means of obtaining more accurate solutions. For hyperbolic equations it could be viewed as a type of hybrid method which combines the desirable error propagation

properties of a less accurate method with the smaller local errors of a more accurate method. For example, the Glimm-Chorin random choice method does an excellent job of resolving shocks, but it is at best first order accurate. Hence the solution obtained may benefit from smoothing followed by a deferred correction with some second order method. Another example is upwind differencing, which, because it is monotone, always converges (Harten, Hyman, and Lax (1976)) to the physically relevant solution of a scalar PDE. Since it is first order, it could also benefit from a deferred correction. It can be shown that if one combines two conservative methods in this way, then the resulting deferred correction method is conservative.

7.1. Test problems. Three hyperbolic conservation laws were used with initial conditions chosen so that the analytical solution is known.

The first problem is

$$\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} u = 0$$

with initial condition

$$u(x, 0) = \begin{cases} 1 , & 0 < x < .3 \\ 0 , & .3 < x < 1 \end{cases}$$

and periodic boundary condition $u(1, t) = u(0, t)$. The analytic solution is a periodic square wave travelling to the right at unit velocity.

The general nonlinear scalar conservation law has the form

$$\frac{\partial}{\partial t} u + \frac{\partial}{\partial x} f(u) = 0 .$$

For initial condition $u(x, 0) = g(x)$ this can be solved via the hodograph transformation to yield the implicit analytical solution

$$u = g(x - f'(u)t)$$

for $u = u(x, t)$. The solution to the differential equation ceases to exist at points where u is about to become multiple-valued. But a weak solution exists having a shock discontinuity originating at such points. The speed $S = S(t)$ of such a shock satisfies the Rankine-Hugoniot condition

$$S = \frac{f(u_R) - f(u_L)}{u_R - u_L}$$

where $u_L = u(S(t)-, t)$ and $u_R = u(S(t)+, t)$. However not all discontinuities satisfying this condition are realizable from continuous initial conditions. Imagine the jump discontinuity $u_R - u_L$ being split into $u - u_L$ and $u_R - u$ where u is strictly between u_L and u_R. In order for these two shocks not to separate, we must have

$$\frac{f(u) - f(u_L)}{u - u_L} \geq \frac{f(u_R) - f(u)}{u_R - u} .$$

This inequality, which must hold for all u strictly between u_L and u_R, is known as the entropy condition.

The choice $f(u) = \frac{1}{2} u^2$ corresponds to the inviscid Burgers equation, whose solution satisfies

$$u = g(x - ut) .$$

The initial condition $u(x, 0) = g(x)$ can be chosen to give interesting analytic solutions. For example, one can arrange to have two shocks develop spontaneously and then later merge into a single shock. We used the initial condition $u(x, 0) = -\arctan x$ and imposed boundary conditions at $x = \pm 1$ obtained from the analytic solution which satisfies $u = -\arctan(x - ut)$ and

$$0 < u < \pi/2 \quad \text{for } x < 0 ,$$

$$u = 0 \quad \text{for } x = 0 ,$$

$$-\pi/2 < u < 0 \quad \text{for } x > 0 .$$

For $t > 1$ there is a stationary shock discontinuity at $x = 0$.

For a system of M conservation laws

$$\frac{\partial}{\partial t}\,\underline{u} + \frac{\partial}{\partial x}\,\underline{f}(\underline{u}) = \underline{0}$$

one might be able to obtain a solution to the Riemann problem, which has the initial conditions

$$\underline{u}(x,\ 0) = \begin{cases} \underline{u}_0, & x < x_0 \ , \\ \underline{u}_M, & x > x_0 \ . \end{cases}$$

For this problem we hypothesize a piecewise constant solution with values $\underline{u}_0, \underline{u}_1, \ldots, \underline{u}_M$ separated by M discontinuities at $x = x_0 + S_m t$, $m = 1(1)M$. The weak form of the PDE reduces to the Rankine-Hugoniot conditions at the discontinuities:

$$S_m(\underline{u}_m - \underline{u}_{m-1}) = \underline{f}(\underline{u}_m) - \underline{f}(\underline{u}_{m-1}),\ m = 1(1)M \ .$$

This represents M^2 nonlinear algebraic equations for the M^2 unknowns $\underline{u}_1, \underline{u}_2, \ldots, \underline{u}_{M-1}$, and $S_1, S_2, \ldots, S_M$. Care must be taken to fix up the solution at those discontinuities where the entropy condition is not satisfied.

The equations for the Eulerian specification of one dimensional gas dynamics are

$$\frac{\partial}{\partial t}\begin{pmatrix} \rho \\ \rho u \\ \rho\varepsilon + \frac{1}{2}\rho u^2 \end{pmatrix} + \frac{\partial}{\partial x}\begin{pmatrix} \rho u \\ \rho u^2 + p \\ u(\rho\varepsilon + \frac{1}{2}\rho u^2 + p) \end{pmatrix}$$

where ρ is density, u is velocity, p is pressure, and ε is internal energy per unit mass. We assume a perfect gas for which the equation of state is

$$p = (\gamma - 1)\rho\varepsilon \ .$$

The specific heat ratio γ is chosen to be 7/5, which is its value for diatomic gases such as air. It is sometimes convenient to use conservation variables

$$\underline{u} = \begin{pmatrix} \rho \\ m \\ e \end{pmatrix} .$$

where m is momentum and e is energy per unit volume. If the fourth variable is eliminated by the equation of state, then

$$\underline{f}(\underline{u}) = \begin{pmatrix} m \\ (\gamma - 1)e + \frac{1}{2}(3 - \gamma)m^2/\rho \\ \gamma me/\rho - \frac{1}{2}(\gamma - 1)m^3/\rho^2 \end{pmatrix} .$$

The analytic solution of the Riemann problem for these equations reduces to a single nonlinear equation, which can be solved by the Godunov iteration. At discontinuities where the entropy condition is violated, one must use a rarefaction wave solution instead. The initial conditions used are those of Sod (1977):

$$\text{at } t = 0 \text{ we have } \rho = 1,\ u = 0,\ p = 1 \text{ for } 0 \le x < \frac{1}{2} \text{ and}$$

$$\rho = 1/8,\ u = 0,\ p = 1/10 \text{ for } \frac{1}{2} < x \le 1 .$$

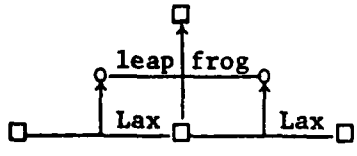
7.2. Two-step Lax-Wendroff method. For the problem $u_t + f(u)_x = 0$ this method consists of a step of Lax's method

$$\frac{1}{k/2}\left(U^{n+1/2}_{j+1/2} - \mu_x U^n_{j+1/2}\right) + \frac{1}{h}\delta_x f(U^n_{j+1/2}) = 0$$

followed by a leapfrog step

$$\frac{1}{k}\delta_t U^{n+1/2}_j + \frac{1}{h}\delta_x f(U^{n+1/2}_j) = 0 ,$$

best illustrated schematically by



For an arbitrary function v the Lax-Wendroff finite difference operator ϕ yields

$$\phi(\Delta v)_j^{n+1} = (v_t + f(v)_x)_j^{n+1/2} - \frac{k}{2}\,(f'(v)(v_t + f(v)_x))_x\Big|_j^{n+1/2} + O(h^2 + k^2)\ .$$

With $v = u$ we get the local error

$$\phi(\Delta u)_j^{n+1} = h^2\,\Psi + O(h^3)$$

where

$$\Psi := \frac{1}{24}\,(cu_{tt} - u_{xx})_t - \frac{1}{8}\,(f'(u)(cu_{tt} - u_{xx}))_x$$

and

$$c := k/h\ .$$

The global error satisfies

$$U_j^n - u_j^n = h^2 e_j^n + O(h^3)$$

where $e = e(x, t)$ solves

$$e_t + (F'(u)e)_x = -\Psi\ .$$

Only the expansion for $\phi(\Delta v)$ is needed for developing deferred correction error estimators.

We managed to obtain fairly close agreement with the numerical results of Sod (1977) for the Lax-Wendroff solution of the Riemann problem. For the time increment he chose

$$k = 0.9\ \max(|u| + c)h\ .$$

For the analytic solution the maximum is attained in Region 4 of Figure 3 of Sod. From $c = \sqrt{\gamma p/\rho}$ and Table II of Sod we get $k \doteq 0.411h$. The value of t in Figures 4 through 15 is not given, but it is clear that the shock position is very close to .75. From equation (15) and Table II the shock speed can be calculated and we get $t/k = 35$. For the Lax-Wendroff method an artificial viscosity term was added. In our test we also ran the problem without artificial viscosity, and in spite of greater oscillations, the average (ℓ_1 norm) error at $t = 35k$ was reduced by 10% for the density,

24% for the pressure, 43% for the velocity, and 16% for the internal energy with similar reductions at earlier time levels. (Note: the column labelled "e" in Table II should be labelled "ε" and the figures labelled "ENERGY" should be labelled "INTERNAL ENERGY.") This may reflect a shortcoming of the ℓ_1 norm, which does not sufficiently penalize oscillations. It has been suggested that the ℓ_2 norm may be more appropriate because convergence results have been obtained for this norm.

For the nonlinear scalar test problem a variable mesh is certainly appropriate, and for this reason we sought a second order accurate variable-mesh generalization of the Lax-Wendroff method. In order to avoid problems near boundaries and to limit computational costs, it was required that $U_{j+1/2}^{n+1/2}$ be a linear combination of U_j^n, $f(U_j^n)$, U_{j+1}^n, and $f(U_{j+1}^n)$ and that U_j^{n+1} be a linear combination of U_j^n, $f(U_j^n)$, $U_{j\pm1}^n$, $f(U_{j\pm1}^n)$, and $f(U_{j\pm1/2}^{n+1/2})$. The method

$$U_{j+1/2}^n = \frac{1}{x_{j+1} - x_j} \left[(x_{j+1} - x_{j+1/2})\, U_j^n + (x_{j+1/2} - x_j)\, U_{j+1}^n \right.$$

$$\left. - \frac{k}{2} \left(F(U_{j+1}^n) - F(U_j^n) \right) \right]$$

$$U_j^{n+1} = U_j^n - \frac{k}{x_{j+1/2} - x_{j-1/2}} \left(F(U_{j+1/2}^{n+1/2}) - F(U_{j-1/2}^{n-1/2}) \right)$$

satisfies these restrictions provided that the whole-numbered meshpoints satisfy

$$x_j = \frac{1}{2} \left(x_{j-1/2} + x_{j+1/2} \right) ;$$

otherwise, there is no second order formula. Given the whole-numbered meshpoints, it is always possible to solve for the half-numbered meshpoints but it may not be possible to force them to lie between the appropriate whole-numbered meshpoints. It would be better to define the half-numbered

meshpoints to be midway between the whole-numbered meshpoints and to define new whole-numbered meshpoints to be midway between the half-numbered meshpoints.

It is well known that without enough artificial viscosity Lax-Wendroff produces oscillations next to a shock. For a bounded stationary numerical solution to $u_t + f(u)_x = 0$ where $f(u)$ is strictly convex or concave one can show that to a first approximation the error decreases geometrically with alternating sign as one moves away from the stationary shock. In section 5 we indicated that errors can be better estimated if they are smooth. A numerical experiment was performed for the Lax-Wendroff method without artificial viscosity applied to the inviscid Burgers equation with $h = 0.1$ and $k = 0.05$. The table that follows contains values for

$$\text{error in } U_j^n ,$$

$$\text{error in } \frac{1}{4} U_{j-1}^n + \frac{1}{2} U_j^n + \frac{1}{4} U_{j+1}^n ,$$

$$\text{error in } \frac{1}{4} (U_{j-1/2}^{n-1/2} + U_{j+1/2}^{n-1/2} + U_{j-1/2}^{n+1/2} + U_{j+1/2}^{n+1/2}) .$$

The $\frac{1}{4}, \frac{1}{2}, \frac{1}{4}$ smoothing is exact for linear polynomials and removes alternating errors of constant magnitude. The second smoothing is suggested by the simple analysis mentioned previously. If we look at the errors in the worst case, it is clear that the second smoothing is better than the first smoothing which in turn is better than not smoothing at all. Similar results were obtained for $h = 0.05$ and $k = 0.025$.

t \ x	0.1	0.2	0.4	0.8
	.0134	.0039	-.0017	-.0003
0.70	.0320	.0217	.0043	.0012
	.0303	.0207	.0046	.0012
	.0228	-.0067	-.0027	-.0003
0.85	.0742	.0197	.0018	.0009
	.0701	.0195	.0022	.0010
	.0008	-.0174	-.0026	-.0002
0.95	.0995	.0052	.0011	.0008
	.0927	.0078	.0013	.0008
	-.0301	-.0194	-.0024	-.0002
1.00	.1033	-.0055	.0010	.0007
	.0948	-.0001	.0011	.0008
	-.0761	-.0163	-.0024	-.0002
1.05	.0997	-.0173	.0011	.0007
	.0892	-.0077	.0010	.0007
	-.2006	.0097	-.0025	-.0002
1.15	.0763	-.0387	.0006	.0006
	.0604	-.0164	.0007	.0006
	-.3692	.0759	-.0007	-.0001
1.30	.0454	-.0541	-.0024	.0005
	.0166	-.0083	-.0005	.0005

In the remainder of this subsection we consider local error estimators $\psi = \phi - \bar{\phi}$ for the Lax-Wendroff method ϕ. We begin by restricting ourselves to operators $\bar{\phi}$ involving only two time-levels n and n+1. Freely available are values of f at U_j^n, $U_{j+1/2}^{n+1/2}$, and U_j^{n+1} for any j. The order of $\bar{\phi}$ cannot exceed the order of $\bar{\phi}$ in time, which can be determined by sending $h \to 0$. Thus we can begin our search for $\bar{\phi}$ by considering discrete-time continuous-space operators, which are little more than ODE methods for IVPs. The Lax-Wendroff method becomes

$$U^{n+1} = U^n + k\, F(U^n + \frac{k}{2} F(U^n))$$

where F denotes the operator $-f(\cdot)_x$ and $U^n = U^n(x)$. With the values of F at U^n, $U^n + \frac{k}{2} F(U^n)$, and U^{n+1} it is not possible to construct a third order implicit Runge-Kutta method $\bar{\phi}$. If we permit one additional

evaluation of F, third order and even fourth order methods $\bar{\phi}$ can be found. However, it is not possible to simultaneously satisfy the accuracy condition

$$\bar{\phi}(\Delta u) = O(k^4)$$

and the contractivity condition

$$\bar{\phi}(\Delta v) = \phi(\Delta v) + O(k^2)$$

for arbitrary functions v, cf. Skeel (1980, p. 35). This rules out the possibility of a fourth order local error estimator ψ. For example, $\bar{\phi}$ given by

$$U^{n+1} = U^n + \frac{k}{6} F(U^n) + \frac{2k}{3} F(\frac{1}{2} U^n + \frac{1}{2} U^{n+1} + \frac{k}{8} F(U^n) - \frac{k}{8} F(U^{n+1})) + \frac{k}{6} F(U^{n+1})$$

is fourth order, but

$$\bar{\phi}(\Delta v)^{n+1} = (v_t - F(v))^{n+1/2} + O(k^2)$$

differs so much from

$$\phi(\Delta v)^{n+1} = (1 + \frac{k}{2} F'(v))(v_t - F(v))\Big|^{n+1/2} + O(k^2)$$

that $\psi = \phi - \bar{\phi}$ yields only a third order error estimate. A reasonably compact extension of this operator to discrete time and discrete space is given by the fourth order scheme

$$\frac{1}{k} \delta_t (1 + \frac{1}{12} \delta_x^2) U_j^{n+1/2} + \frac{1}{h} \delta_x (\frac{2}{3} + \frac{1}{3} \mu_x \mu_t) f(U_j^{n+1/2}) = 0$$

where

$$U_{j+1/2}^{n+1/2} := \mu_t \mu_x (1 - \frac{1}{8} \delta_x^2) U_{j+1/2}^{n+1/2} + \frac{k}{8} \mu_t \delta_x (1 - \frac{1}{24} \delta_x^2) f(U_{j+1/2}^{n+1/2}) .$$

If two additional evaluations of f are permitted, one can obtain a fourth order estimate by using a scheme of Abarbanel, Gottlieb, and Turkel described by Morton (1976), which generalizes the classical Runge-Kutta formula to $u_t + f(u)_x = 0$:

$$U_{j+1/2}^{n+1/2} = \mu_x U_{j+1/2}^{n} - \frac{k}{2h} \delta_x f(U_{j+1/2}^{n}) ,$$

$$U_{j}^{n+1/2} = (1 - \frac{1}{8} \delta_x^2) U_{j}^{n} - \frac{k}{2h} \delta_x f(U_{j}^{n+1/2}) ,$$

$$U_{j+1/2}^{n+1} = \mu_x (1 - \frac{1}{8} \delta_x^2) U_{j+1/2}^{n} - \frac{k}{h} \delta_x [f(U_{j+1/2}^{n+1/2}) + \frac{1}{8} \delta_x^2 f(U_{j+1/2}^{n})] ,$$

$$U_{j}^{n+1} = U_{j}^{n} - \frac{k}{h} \delta_x [\frac{1}{6} f(U_{j}^{n+1}) + \frac{1}{3} (\mu_x + 1 - \frac{1}{4} \delta_x^2) f(U_{j}^{n+1/2}) + \frac{1}{6} \mu_x (1 - \frac{1}{8} \delta_x^2) f(U_{j}^{n})] .$$

We compared this method to Lax-Wendroff for the gas dynamics problem without artificial viscosity. The average error for the fourth order method at t = 35k was greater by 8% for the density, 20% for the pressure, 64% for the velocity, and 14% for the internal energy. Also it produced a 54% overshoot in the velocity compared to 33% for Lax-Wendroff.

If one is willing to use a three-level operator $\bar{\phi}$, then one can avoid additional evaluations of f by using

$$\frac{1}{k} \mu_t \delta_t (1 + \frac{1}{6} \delta_x^2) U_{j}^{n+1} + \frac{1}{h} \mu_x \delta_x (1 + \frac{1}{6} \delta_t^2) f(U_{j}^{n+1}) = 0 .$$

The prospects of obtaining good error estimates with these operators $\bar{\phi}$ did not seem very good, and so we turned our attention to estimating errors of first order methods.

7.3. Forward-time centered-space differencing. This method is known to be unstable for the linear model problem, but it was accidentally tested due to erroneous programming of Lax's method. The linear test problem was solved with h = 1/10 and k = 1/60, and a deferred correction solution was computed using Lax-Wendroff. Below we have the

error in the FTCS solution

error estimate from deferred correction

for all the spatial meshpoints at t = 30k:

.30 -.25 -.69 .00 1.06 .00 .13 -.87 .19 .12

.31 -.11 -.46 -.08 .47 -.08 .44 -.47 .09 -.08

Out of interest we also computed the Lax-Wendroff solution, and below we have the

error in deferred correction solution

error in Lax-Wendroff solution

.00 -.14 -.22 .08 .59 .08 -.31 -.40 .11 .21

.12 -.16 -.34 .07 .74 -.05 -.13 -.60 .20 .15

In sum, deferred correction performed well in this instance.

7.4. Lax's method. This is a highly dissipative first order scheme given by

$$U_j^{n+1} = \frac{1}{2}(U_{j-1}^n + U_{j+1}^n) - \frac{k}{h}\mu_x\delta_x f(U_j^n) .$$

For the linear test problem with h = 1/10 and k = 1/60 we tried for $\bar{\phi}$ both Lax-Wendroff and Lax-Wendroff with diffusive and antidiffusive fluxes as described by Sod (1977). In both cases the results were disastrous probably due to the fact that the numerical solution separates into two uncoupled solutions, one with j+n even and the other with j+n odd. Lindberg (1976) encountered the same problem with the leapfrog scheme. Out of interest we also computed solutions directly with each of the $\bar{\phi}$ schemes and the average error was 17% less with diffusive and antidiffusive fluxes.

We repeated the experiment with leapfrog for $\bar{\phi}$ and with k = 1/50. The deferred correction solution was nearly the same as the Lax solution, and thus was useless for global error estimation.

In order to avoid separated solutions we considered instead a two-step Lax scheme in which we first compute $U_{j\pm1/2}^{n+1/2}$ using the Lax method and then compute U_j^{n+1} from these two values again using the Lax method. With $k = 1/60$ and $\bar{\phi}$ = Lax-Wendroff we computed a deferred correction solution. Below we have the

error in the Lax solution

error estimate from deferred correction

for all the meshpoints at $t = 20k$:

.23	.26	.30	.35	-.63	-.63	.34	.30	.25	.23
.14	.07	-.02	-.12	-.15	-.14	-.07	.02	.12	.15

These are not very good results. The same combination for ϕ and $\bar{\phi}$ was also tested on Burgers' equation with $h = 1/10$ and $k = 1/50$. The results were much better. Below we give the

time level

average error in Lax solution

average error estimate from deferred correction

1	3	7	15	30
.0014	.0041	.0099	.0233	.0572
.0013	.0039	.0090	.0194	.0417

We also compared the deferred correction solution to the Lax-Wendroff solution and found that by the 30th time level the average error of the latter was smaller by a factor of 14 but that spurious oscillations were less pronounced for the former. The deferred correction error was smaller than the Lax error by a factor of 4.

7.5. Upwind differencing. For Burgers' equation we tested the conservative upwind scheme

$$\frac{1}{k}\,(u_j^{n+1} - u_j^n) + \frac{1}{h}\,(g_{j+1/2}^n - g_{j-1/2}^n) = 0$$

where (omitting n)

$$g_{j+1/2} = \frac{1}{2}\,f(u_{j+1}) + \frac{1}{2}\,f(u_j) - \frac{1}{2}\,\mathrm{sgn}(f'(\tfrac{1}{2}\,u_{j+1} + \tfrac{1}{2}\,u_j))(f(u_{j+1}) - f(u_j))\,.$$

If $f'(u)$ is positive for u close to u_j^n, this has the effect of defining u_j^{n+1} in terms of u_{j-1}^n and u_j^n, which is desirable because the characteristic passing through (x_j, t^{n+1}) would pass between (x_{j-1}, t^n) and (x_j, t^n). If $f'(u)$ is negative, then u_j^{n+1} is given in terms of u_j^n and u_{j+1}^n. With $h = 1/10$, $k = 1/50$, and $\bar{\phi}$ = Lax-Wendroff we computed a deferred correction solution. Below we give the

time level

average error in upwind solution

average error estimate from deferred correction

1	3	7	15	30	35
.0003	.0010	.0025	.0054	.0106	.0112
.0004	.0011	.0025	.0053	.0089	.0092

Out of interest we also computed the Lax-Wendroff solution, and below we have the

time level

average error in deferred correction solution

average error in Lax-Wendroff solution

1	3	7	15	30	35
.00002	.00007	.00017	.00051	.00238	.00335
.00002	.00007	.00017	.00046	.00164	.00246

For the gas dynamics equation Roache (1975, p. 237) defines upwind differencing by

$$\underline{U}_j^{n+1} = \underline{U}_j^n - \text{sgn} \cdot \frac{k}{h} (\underline{G}_j^n - \underline{G}_{j-\text{sgn}}^n) - \frac{k}{2h} (\underline{S}_{j+1}^n - \underline{S}_{j-1}^n) ,$$

where $\underline{G} = (\rho u, \rho u^2, u(e+p))^T$, $\underline{S} = (0, p, 0)^T$ and $\text{sgn} = \text{sgn}(u_j^n)$. (The equations given by Sod (1977, p. 30) appear to be erroneous.) The numerical results obtained by Sod were not good. There was a nonphysical shock where there should have been an expansion wave. Our results were somewhat similar except that there were also oscillations present. The stability condition $\sigma \leq 1$ given by Sod differs from the very stringent condition given by Roache. This latter condition is violated in the experiments.

Because of the failure of this upwind differencing scheme for the system of three hyperbolic equations, we tried to extend the scalar scheme used for Burgers' equation to a system of conservation laws. A systematic way of doing this was discussed in the introduction to this section. For upwind differencing it involves using $\text{sgn}(\underline{f}'(\underline{u}))$ where if $\underline{f}'(\underline{u}) = Q\ \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_M)Q^{-1}$, then $\text{sgn}(\underline{f}'(\underline{u})) = Q\ \text{diag}(\text{sgn}(\lambda_1), \text{sgn}(\lambda_2), \ldots, \text{sgn}(\lambda_M))Q^{-1}$, which is independent of Q. For the gas dynamics equations it is most convenient to use the variables ρ, u, and $c = \sqrt{\gamma(\gamma-1)\varepsilon}$, and after substitution into $\underline{f}'(\underline{u})$ we have

$$\underline{f}'(\underline{u}) = \begin{pmatrix} 0 & 1 & 0 \\ -\frac{(3-\gamma)u^2}{2} & (3-\gamma)u & \gamma-1 \\ -\frac{uc^2}{\gamma-1} - \frac{(2-\gamma)u^3}{2} & \frac{c^2}{\gamma-1} + \frac{(3-2\gamma)u^2}{2} & \gamma u \end{pmatrix}$$

$$= Q\ \text{diag}(u, u+c, u-c)Q^{-1}$$

where

$$Q = \begin{pmatrix} 1 & 1 & 1 \\ u & u + c & u - c \\ \frac{u^2}{2} & \frac{u^2}{2} + uc + \frac{c^2}{\gamma-1} & \frac{u^2}{2} - uc + \frac{c^2}{\gamma-1} \end{pmatrix}$$

and

$$Q^{-1} = \frac{\gamma-1}{c^2} \begin{pmatrix} \frac{c^2}{\gamma-1} - \frac{u^2}{2} & u & -1 \\ \frac{u^2}{4} - \frac{uc}{2(\gamma-1)} & -\frac{u}{2} + \frac{c}{2(\gamma-1)} & \frac{1}{2} \\ \frac{u^2}{4} + \frac{uc}{2(\gamma-1)} & -\frac{u}{2} - \frac{c}{2(\gamma-1)} & \frac{1}{2} \end{pmatrix} .$$

The following may be useful in organizing the computation:

$$\underline{u} = Q \cdot \frac{\rho}{\gamma} \begin{pmatrix} \gamma-1 \\ 1/2 \\ 1/2 \end{pmatrix}$$

and

$$\underline{f}(\underline{u}) = \underline{f}'(\underline{u})\underline{u} .$$

Remark. The nonconservative upwind differencing given by Roache would be in line with the above approach if we had used the splitting $\underline{f} = \underline{G} + \underline{S}$ where $\underline{G} = \frac{\gamma-1}{\gamma} (\rho u, \rho u^2, \frac{1}{2} \rho u^3)$ as long as $|u| < c$.

The matrix upwind differencing scheme described in the paragraph before the remark was tested on the Riemann problem, and it produced a very nice solution. With $\bar{\phi}$ = Lax-Wendroff (without artificial viscosity) we computed a deferred correction solution. Compared to the original matrix upwind solution, the deferred correction solution had average errors at t = 35k which were worse by more than a factor of 2 for all variables. In addition a spurious rarefaction shock and compression shock appeared between the rarefaction expansion and the contact discontinuity.

A compact second order but implicit scheme is the box scheme

$$\frac{1}{k}\,\delta_t\mu_x\,U^{n+1/2}_{j+1/2} + \frac{1}{h}\,\delta_x\mu_t\,f(U^{n+1/2}_{j+1/2}) = 0\ .$$

Whether this should be used to determine U^{n+1}_{j+1} from U^{n+1}_{j} or vice versa depends on the direction of the characteristics. A conservative upwind box scheme is given by

$$U^{n+1}_j = U^n_j - \frac{k}{h}\,\delta_x\,g^{n+1/2}_j$$

where

$$g^{n+1/2}_{j+1/2} = \frac{h}{4k}\,\delta_t\delta_x\,U^{n+1/2}_{j+1/2} + \mu_t\mu_x\,f(U^{n+1/2}_{j+1/2})$$

$$- \frac{1}{2}\,\mathrm{sgn}\,f'(\mu_t\mu_x\,U^{n+1/2}_{j+1/2})\left[\frac{h}{k}\,\delta_t\mu_x\,U^{n+1/2}_{j+1/2}\right.$$

$$\left. + \mu_t\delta_x\,f(U^{n+1/2}_{j+1/2})\right]\ .$$

If we were to use this as a solution method, one could apply this iteratively beginning with $U^{n+1}_j = U^n_j$ for all j. Doing one iteration is equivalent to our first order upwind scheme, and doing two iterations yields second order accuracy. This scheme extends to a system as we have described. We tested this scheme as a solution method for the Riemann problem and found that with two iterations per time level the average error in the solution variables was comparable to that of the first order upwind differencing. With three iterations per time level the error was decidedly worse.

A deferred correction solution was computed from the upwind solution of the Riemann problem using $\bar{\phi}$ = upwind box with 2 iterations. (We discovered that $\bar{\phi}(U)$ would remain unchanged if we had instead used $\bar{\phi}$ = implicit upwind box.) Below we have given for three different time levels the

average errors in matrix upwind solution

average error estimates from deferred correction

time level	ρ	p	u	ε
5	.0052	.0059	.0107	.0175
	.0024	.0032	.0082	.0127
15	.0088	.0082	.0152	.0294
	.0049	.0052	.0113	.0193
35	.0125	.0108	.0200	.0440
	.0075	.0071	.0139	.0269

We also computed the solution in two other ways, and below we have the average errors at various time levels for

deferred correction with $\bar{\phi}$ = upwind box with 2 iterations

deferred correction with $\bar{\phi}$ = upwind box with 3 iterations

Lax-Wendroff (without artificial viscosity)

time level	ρ	p	u	ε
5	.0043	.0044	.0087	.0118
	.0040	.0040	.0084	.0108
	.0064	.0069	.0150	.0181
15	.0056	.0052	.0105	.0192
	.0051	.0046	.0093	.0174
	.0079	.0071	.0156	.0242
35	.0073	.0062	.0148	.0314
	.0070	.0059	.0143	.0299
	.0084	.0064	.0127	.0279

References

S. Abarbanel, D. Gottlieb, and E. Turkel, "Difference schemes with fourth order accuracy for hyperbolic equations," Modern Developments in Fluid Dynamics, J. Rom ed., SIAM.

J. Christiansen and R. D. Russell (1979), "Deferred corrections using nonsymmetric end formulas," submitted to Num. Math.

B. Epstein and D. Hicks (1979), "Computational experiments on two error estimation procedures for ordinary differential equations," AFWL-TR-78-119, Air Force Weapons Lab, Kirtland AFB, NM.

L. Fox (1947), "Some improvements in the use of relaxation methods for the solution of ordinary and partial differential equations," Proc. Roy. Soc. London Ser. A., 190, 31-59.

W. Hackbusch (1977), "Extrapolation applied to certain discretization methods solving the initial value problem for hyperbolic differential equations," Num. Math. 28, I21-142.

A. Harten, J. M. Hyman, and P. D. Lax (1976), "On finite-difference approximations and entropy conditions for shocks," COO-3077-106, Courant Institute of Mathematical Sciences, New York University.

H. B. Keller (1976), Numerical Solution of Two-Point Boundary-Value Problems, Regional Conference Series in Appl. Math., SIAM.

P. D. Lax and B. Wendroff (1960), "Systems of conservation laws," Comm. Pure Appl. Math. 13, 217.

M. Lentini and V. Pereyra (1977), "An adaptive finite difference solver for nonlinear two-point boundary problems with mild boundary layers," SIAM J. Numer. Anal. 14, 91-111.

B. Lindberg (1976), "Error estimation and iterative improvement for the numerical solution of operator equations," UIUCDCS-R-76-820, Dept. of Computer Sci., Univ. of Illinois at Urbana-Champaign.

B. Lindberg (1980), "Error estimation and iterative improvement for discretization algorithms," to appear in BIT.

K. W. Morton (1976), "Initial-value problems by finite difference and other methods," The State of the Art in Numerical Analysis, D. Jacobs ed., Academic Press.

P. J. Roache (1975), Computational Fluid Dynamics, Hermosa, Albuquerque.

R. D. Skeel (1980), "A theoretical framework for proving accuracy results for deferred corrections," submitted to a technical journal. (Also UIUCDCS-F-80-892, Dept. of Computer Sci., Univ. of Illinois at Urbana-Champaign.)

R. D. Skeel and L. W. Jackson (1979), "The stability of variable-step Nordsieck methods," manuscript, to appear in SIAM J. Numer. Anal.

G. A. Sod (1977), "A survey of numerical methods for compressible fluids," COO-3077-145, Courant Institute of Mathematical Sciences, New York University. See also J. Comp. Phys. 27, 1-31 (1978).

M. Spijker (1971), "On the structure of error estimates for finite difference methods," Num. Math. 18, 73-100.

H. J. Stetter (1971), "Local estimation of the global discretization error," SIAM J. Numer. Anal. 8, 512-523.

H. J. Stetter (1978), "The defect correction principle and discretization methods," Num. Math. 29, 425-443.

F. Stummel (1975), "Biconvergence, bistability, and consistency of one-step methods for the numerical solution of initial value problems in ordinary differential equations," Topics in Numerical Analysis II, J.J.M. Miller ed., Academic Press, London, 197-211.

P. N. Swarztrauber and R. A. Sweet (1979), "ALGORITHM 541: Efficient Fortran subprograms for the solution of separable elliptic partial differential equations," ACM Trans. Math. Software 5, 3, 352-364.

II. GAUSSIAN ELIMINATION AND NUMERICAL INSTABILITY

The solution of linear systems of equations is basic to the numerical solution of many problems, and yet this subproblem has not been treated in an entirely satisfactory way. Stewart (1973) states that "In spite of intense theoretical investigation, there is no satisfactory algorithm for scaling a general matrix." A theoretical solution to the scaling problem was discovered by Skeel (1979). This was one of the results of a study of the implications of a stability concept which is more appropriate for sparse systems. This investigation was stimulated by a report of Gear (1975).

1. Scaling for Numerical Stability in Gaussian Elimination (Skeel; Ortman)

A paper with this title was authored by Skeel (1979). The abstract follows:

> Roundoff error in the solution of linear algebraic systems is studied using a more realistic notion of what it means to perturb a problem, namely, that each datum is subject to a relatively small change. This is particularly appropriate for sparse linear systems. The condition number is determined for this approach. The effect of scaling on the stability of Gaussian elimination is studied, and it is discovered that the proper way to scale a system depends on the right-hand side. However, if only the norm of the error is of concern, then there is a good way to scale that does not depend on the right-hand side.

The table of contents is as follows:

Measurements of certain quantities introduced in this paper were made for the LINPACK SGEFA test problems. Unfortunately for our purposes these problems were not representative of those likely to occur in practice, but rather there were a number of unusual problems designed to test the logic of SGEFA. For all but one of the problems the backward error

$$\eta := \max \frac{|b - A\hat{x}|}{|A|\,|\hat{x}|}$$

was less than the unit roundoff error u even though the ill scaling ratio

$$\sigma_R(A, x) := \frac{\max |A| |x|}{\min |A| |x|}$$

was as high as $1.6_{10}29$. For problem 7 the backward error was 5 units of roundoff error and the ill-scaling ratio was $3.0_{10}7$.

2. Iterative Improvement Implies Numerical Stability for Gaussian Elimination (Skeel)

A paper with this title was authored by Skeel (1980). The abstract follows:

Because of scaling problems, Gaussian elimination with pivoting is not always as accurate as one might reasonably expect. It is shown that even a single iteration of iterative refinement in single precision is enough to make Gaussian elimination stable in a very strong sense. Also, it is shown that without iterative refinement row pivoting is inferior to column pivoting in situations where the norm of the residual is important.

The table of contents is

3. Effect of Equilibration on Residual Size for Partial Pivoting (Skeel)

A paper with this title was authored by Skeel (198x). The abstract follows:

It is shown that column pivoting with row equilibration satisfies the same type of error bound as does row pivoting without scaling and that row pivoting with column equilibration satisfies the same type of bound as does column pivoting without scaling. An interesting consequence for row pivoting is that column equilibration is sufficient to ensure that the norm of the residual is reasonably small.

The table of contents is

References

C. W. Gear (1975), "Numerical errors in sparse linear equations," UIUCDCS-F-75-885, Dept of Computer Sci., Univ. of Illinois at Urbana-Champaign.

R. D. Skeel (1979), "Scaling for numerical stability in Gaussian elimination," J. ACM 26, 3, 494-526.

R. D. Skeel (1980), "Iterative improvement implies numerical stability for Gaussian elimination," Math. Comp. 34, 151, to appear.

R. D. Skeel (198x), "Effect of equilibration on residual size for partial pivoting," SIAM J. Numer. Anal., to appear. (Also, UIUCDCS-F-79-891, Dept. of Computer Sci., Univ. of Illinois.)

G. W. Stewart (1973), Introduction to Matrix Computations, Academic Press, New York.

III. NUMERICAL SOLUTION OF ORDINARY DIFFERENTIAL EQUATIONS

Systems of ODEs occur naturally in many applications and also arise from the spatial discretization of time-dependent partial differential equations. The most effective numerical methods for solving initial value problems in ordinary differential equations have been multistep methods. Work in this area is reported here.

1. Equivalent Forms of Multistep Formulas (Skeel)

The Adams-Bashforth-Moulton and the backward differentiation formulas are very popular. Improved stability properties are possible with other formulas, but they are not used for several reasons. One of the difficulties concerns storage requirements. It is shown in Skeel (1979) how to formulate the corrector formula and how to select a predictor so that any multistep method can be implemented as cheaply as the two popular ones. The abstract of this paper follows:

> For uniform meshes it is shown that any linear k-step formula can be formulated so that only k values need to be saved between steps. By saving additional m values it is possible to construct local polynomial approximations of degree k + m - 1, which can be used as predictor formulas. Different polynomial bases lead to different equivalent forms of multistep formulas. In particular, local monomial bases yield Nordsieck formulas. An explicit one-to-one correspondence is established between Nordsieck formulas and k-step formulas of order at least k, and a strong equivalence result is proved for all but certain pathological cases. Equivalence is also shown for P(EC)* formulas but not for P(EC)*E formulas.

The table of contents is as follows:

2. The Stability of Variable-Step Nordsieck Methods (Skeel and Jackson)

Work on the stability of interpolatory step changing has been prepared for publication by Skeel and Jackson (198x). The abstract follows:

> Conditions are given for the stability of the Nordsieck formulation of Adams and backward differentiation methods. It is shown that if the stepsize selection function is variation-bounded, then these methods are stable. It is proven that for each method there exist constants a, b, c satisfying $a \leq b < 1 < c$ such that if the stepsize ratio h_{n+1}/h_n satisfies $0 < h_{n+1}/h_n < a$ or if $b < h_{n+1}/h_n < c$ then the method is stable. For k-value methods with $k \leq 8$, tables are given for the values a, b, and c. If the stepsize is kept constant for one or more steps, then methods are more stable in the sense that the intervals defined by a, b, and c are larger. Tables and graphs are given which show how the stability of a method changes as the stepsize is changed less often.

The table of contents is as follows:

3. Blended Linear Multistep Methods (Skeel; Dahlquist-Downs, Vu)

At present the most popular codes for stiff systems of ODEs are based on the backward differentiation formulas. One promising alternative is to use the "blended" multistep formulas (Skeel and Kong (1977)) which attempt to combine the best features of the Adams and the backward differentiation formulas. By definition every autonomous stiff system must have at least one "component" which is nonstiff and the Adams formulas are very good for such equations. There are theoretical reasons for believing that the blended methods are both effective and computationally inexpensive. Limited empirical evidence suggests that the blended formulas may be as good as the backward differentiation formulas for stiff problems, better for nonstiff problems, and much better for stiff oscillatory problems. We have been working on implementing these methods by making modifications to state-of-the-art computer codes for stiff ODEs.

In 1978 from the University of Toronto we acquired STIFF DETEST, which is a testing program for stiff ODE integrators. We attempted to execute the program on an INTERDATA minicomputer, but the object code was too long, and so we wrote our own testing program. An improved version of STIFF DETEST was acquired in 1979, and we have successfully executed it in five and a half hours on a departmental PRIME computer. It seems that the execution time is much too great, and we are looking into this problem.

The blended formulas had been implemented in a structured version of new DIFSUB, which has been distributed to several researchers. We compared the efficiency of this to that of GEAR, Rev. 3 of Hindmarsh (1974) on several test problems. Testing was performed as described by Skeel and Kong (1977) except that we used the CYBER 175 and h_{min} was set to $4ut_f$, where u is the unit roundoff error 2^{-47}.

ε	max order	steps	func evals	backsolves	LU decomps	accurate digits	time
			Numerical Results for Problem 1				
			GEAR				
10^{-3}	4	51	110	76	11	2.9	2.7
10^{-4}	4	69	147	104	14	3.6	3.1
			blended new DIFSUB				
10^{-2}	4	33	98	134	10	3.1	2.6
10^{-3}	7	65	158	236	13	3.8	5.5
			Numerical Results for Problem 2				
			GEAR				
10^{-3}	4	96	211	130	20	2.2	5.2
10^{-4}	5	146	283	190	23	3.4	8.9
			blended new DIFSUB				
10^{-2}	4	65	237	272	25	2.7	7.2
10^{-3}	6	99	283	396	21	3.8	10.0
			Numerical Results for Problem 3				
			GEAR				
10^{-3}	(4)	(1001)	(1385)	(1060)	(54)	(1.5)	(59.0)
	"too much work"; integration stopped at t = 8.1						
			blended new DIFSUB				
10^{-2}	5	139	382	570	16	2.4	19.0
10^{-3}	7	210	520	822	18	3.4	28.0
			Numerical Results for Problem 4				
			GEAR				
10^{-3}	3	111	257	164	23	1.8	6.4
10^{-4}	5	174	358	233	31	3.0	11.0
			blended new DIFSUB				
10^{-2}	(5)	(311)	(1376)	(1782)	(121)	(-1.2)	(44.0)
	"h_{min} too large"; integration stopped at t = 956						
10^{-3}	6	166	503	716	36	2.5	17.0
			Numerical Results for Problem 5				
			GEAR				
10^{-3}	5	67	154	129	6	0.1	5.0
10^{-4}	4	107	240	207	8	0.7	7.2
			blended new DIFSUB				
10^{-2}	6	71	311	380	30	1.6	10.0
10^{-3}	8	79	310	394	28	2.9	12.0

We had planned to implement the blended formulas in GEAR. We made changes to GEAR so as to avoid fixed dimension arrays and local storage. Improvements were made to the calling sequence incorporating some ideas from the user interface standard of Hindmarsh (1978). We also added the initial stepsize selection algorithm of Shampine and Stetter in the source code of RKSW, which is available as a microfiche supplement to Shampine and Wisniewski (1978). We abandoned this effort when Hindmarsh's new code LSODE became available. This code already has the desirable changes, and we are putting the blended formulas into it.

A catalog of better FORTRAN codes for IVPs has been compiled. It has been included in the "Working Papers of the SIGNUM Meeting on Numerical ODEs" and in the June 1979 issue of the SIGNUM Newsletter.

A device has been developed for improving the convergence of the corrector iteration in cases where coefficient in the decomposed matrix is out of date. Suppose we want to solve

$$hp' + \Delta - hf(t, p + \beta\Delta) = 0$$

for Δ given a triangular factorization of $I - r^{-1}h\beta J$ where $J \approx f_y$ and r is the ratio of the current value of $h\beta$ to an old value of $h\beta$. A generalization of the usual correction iteration is given by

$$\Delta_{(m+1)} = \Delta_{(m)} - \omega_m[I - r^{-1}h\beta J]^{-1}\ \text{residual}(\Delta_{(m)})$$

where we have introduced the relaxation factor ω_m instead of using a factor of unity. For the test equation $f(t, y) = \lambda y + g(t)$ with $J = \lambda$ the iteration error $\varepsilon_{(m)} := \Delta_{(m)} - \Delta$ satisfies

$$\varepsilon_{(m+1)} = [r - \mu]^{-1}[r(1 - \omega_m) - (1 - \omega_m r)\mu]\varepsilon_{(m)}$$

where $\mu := \beta h\lambda$. If we do not know the number of iterations beforehand, we could try to optimize the error reduction for each iteration separately

in which case ω_m is the same value ω for each iteration. (If there is at least two iterations, we would optimize ω_1 and ω_2 together and then ω_3 and then ω_4, etc. One might also require 100% error reduction for $\lambda = 0$.) If we restrict λ so that Re $\lambda \leq 0$ then Re $\mu \leq 0$ for the methods of interest. The convergence factor is analytic for Re $\mu \leq 0$, and so by the maximum modulus theorem the maximum value of the convergence factor for all Re h $\lambda \leq 0$ is given by

$$\sup_{-\infty<v<+\infty} |r(1 - \omega) - (1 - \omega r)iv|/|r - iv| .$$

(It may be more appropriate to compute the maximum over all h in the intersection of the left half plane with the absolute stability region of the formula.) The square of the worst case convergence factor is

$$\max_v (r^2(1 - \omega)^2 + (1 - \omega r)^2 v^2)/(r^2 + v^2) .$$

Differentiation with respect to v yields extrema at $v^2 = 0$ and $v^2 = \infty$. Therefore the worst case convergence factor is

$$\max \{|1 - \omega|, |1 - r\omega|\}$$

for which

$$\text{minimum} = \frac{|r - 1|}{r + 1} \text{ at } \omega = \frac{2}{r + 1} .$$

This is smaller by a factor of $1/(r + 1)$ than the convergence factor for $\omega = 1$. Experiments were performed with LSODE for problems 1, 2, and 3, but the results were inconclusive. This device was discovered independently by Chipman (1979).

The Nordsieck implementation of the blended formulas obscures the identity of the predictor formulas that are used, and interest has been expressed in this question. Using the ideas of Skeel (1980, section 2, paragraphs 3, 4, 5), we determined predictor formulas for the k-step blended formulas for k = 1, 2, 3. The predictor for k = 1 is the Euler

formula. For k = 2 it is

$$- y_n + y_{n-1} + \frac{3}{2} hy'_{n-1} - \frac{1}{2} hy'_{n-2}$$

$$+ 6\gamma hJ\{- \frac{1}{2} y_n + \frac{1}{2} y_{n-2} + hy'_{n-1}\} = 0$$

and for k = 3 it is

$$- y_n + y_{n-1} + \frac{23}{12} hy'_{n-1} - \frac{4}{3} hy'_{n-2} + \frac{5}{12} hy'_{n-3}$$

$$+ 10\gamma hJ\{- \frac{23}{30} y_n + \frac{1}{5} y_{n-1} + \frac{9}{10} y_{n-2} - \frac{1}{3} y_{n-3}$$

$$+ \frac{26}{15} hy'_{n-1} - \frac{13}{15} hy'_{n-2} + \frac{2}{15} hy'_{n-3}\}$$

$$+ 64(\gamma hJ)^2\{- \frac{1}{3} y_n - \frac{1}{2} y_{n-1} + y_{n-2} - \frac{1}{6} y_{n-3} + hy'_{n-1}\} = 0 .$$

4. Equivalent Forms of Variable Step Multistep Formulas (Skeel; Vu)

A manuscript with this title is being prepared for publication. Widely available codes such as A. C. Hindmarsh's LSODE use Nordsieck's interpolatory technique to vary the stepsize. It is often stated that this technique does not yield truly variable step formulas. However, we have shown that this is not true in the sense that there exists a formula depending *only* on the meshpoints t_n, t_{n-1},..., t_{n-k} which relates the computed values y_n, y_{n-1},..., y_{n-k} and the derivatives y'_n, y'_{n-1},..., y'_{n-k}. This result may be useful in improving the estimation of local errors in codes that use the interpolatory technique. Other codes, notably EPISODE, are based on "natural" variable step Adams and backward differentiation formulas. The implementation uses the scaled derivatives of the "modifier polynomials" associated with these formulas. We were interested in the result of "blending" these two modifier polynomials, for example, the $AMF^{(3)}$ modifier polynomial of degree 2 at t_n plus $-h_n\gamma_n J_n$ times the $BDF^{(2)}$ modifier polynomial of degree 2 at t_n. After a great deal of algebra we determined the equivalent multistep formula satisfied by the computed solution and derivative values:

$$\{AMF^{(3)}\}_n - h_n\gamma_n J_n\{BDF^{(2)}\}_n$$

$$+ \frac{h_n^3}{(h_n+h_{n-1})h_{n-1}^2}(h_{n-2}\gamma_{n-1}J_{n-1} - h_{n-1}\gamma_n J_n)(I + 3h_{n-2}\gamma_{n-1}J_{n-1})^{-1}\{AMF^{(2)}\}_{n-1} = 0 .$$

This is not a blend of $AMF^{(3)}$ and $BDF^{(2)}$ even for constant J_n unless γ_n are chosen in a manner incompatible with good stability behavior.

References

F. Chipman (1979), "Some experiments with STRIDE," in Working Papers for the 1979 SIGNUM Meeting on Numerical Ordinary Differential Equations, R. D. Skeel ed., UIUCDCS-R-79-963, Dept. of Computer Science, Univ. of Illinois at Urbana-Champaign.

A. C. Hindmarsh (1974), "GEAR, Ordinary differential equation solver," UCID-3001, Rev. 3, Lawrence Livermore Lab, Univ. of California, Livermore.

A. C. Hindmarsh (1978), "A tentative user interface standard for ODEPACK," UCID-17954, Lawrence Livermore Lab, Univ. of California, Livermore.

L. F. Shampine and J. A. Wisniewski (1978), "The variable order Runge-Kutta code RKSW and its performance," SAND78-1347, Sandia Lab, Albuquerque, New Mexico.

R. D. Skeel (1979), "Equivalent forms of multistep formulas," Math. Comp. 33, 148, 1229-1250.

R. D. Skeel and L. W. Jackson (198x), "The stability of variable-step Nordsieck methods," SIAM J. Numer. Anal., to appear. (Also, T.R. 89, Dept. of Computer Sci., Univ. of Toronto.)

R. D. Skeel and A. K. Kong (1977), "Blended linear multistep methods," ACM Trans. Math. Software 3, 4, 326-345.

IV. MULTIGRID METHODS (Van Rosendale; Skeel)

A Ph.D. thesis on this topic was written by Van Rosendale (1980).

The abstract follows:

This thesis is concerned with the use of multi-level methods to solve the linear systems arising from finite element discretizations of elliptic equations. In all, three multi-level methods are considered. The first of these is applicable only to quasi-uniform grids, but is simpler than other algorithms considered in previous theoretical work. The other two algorithms are applicable to both quasi-uniform grids, and locally refined grids, those grids on which the size of the largest and smallest elements may differ by an arbitrarily large factor. All three algorithms are asymptotically optimal, producing good solutions in $O(N)$ operations on a finite element grid with N elements. These asymptotically optimal complexity bounds for the last two algorithms are the first such bounds for multi-level methods on locally refined grids. One of these algorithms achieves this $O(N)$ complexity bound under weaker than expected conditions on the dimensions of the finite element spaces used by the algorithm.

The multi-level convergence results for locally refined grids shown here are based on a new approximation result given in this thesis. This approximation result is of interest for several reasons, the main one being that it is completely local, making no use of global properties such as the regularity of the problem. In consequence, it provides an independent demonstration of the asymptotically optimal complexity of multi-level algorithms on non-convex domains, shown previously by Bank and Dupont. It also permits one to determine explicit upper bounds on the rate of convergence of multi-level methods on irregular finite element grids using only local properties of the finite element space involved.

The table of contents is as follows:

4. Locally Refined Grids
   - 4.1. Introduction
   - 4.2. Notation
   - 4.3. Algorithms
   - 4.4. Complexity
   - 4.5. Interpolation
   - 4.6. Approximation

Reference

J. R. Van Rosendale (1980), "Rapid solution of finite element equations on locally refined grids by multi-level methods," UIUCDCS-R-80-1021, Dept. of Computer Sci., Univ. of Illinois at Urbana-Champaign.

PUBLICATIONS IN TECHNICAL JOURNALS
US AFOSR-75-2854

R. D. Skeel, "Scaling for numerical stability in Gaussian elimination," J. ACM 26, 3 (July 1979), 494-526.

R. D. Skeel, "Equivalent forms of multistep methods," Math. Comp. 33, 148 (October 1979), 1229-1250.

R. D. Skeel, "Iterative refinement implies numerical stability for Gaussian elimination," Math. Comp. 34, 151 (July 1980), to appear.

R. D. Skeel and L. W. Jackson, "The stability of variable-step Nordsieck methods," SIAM J. Numer. Anal., to appear.

R. D. Skeel, "Effect of equilibration on residual size for partial pivoting," SIAM J. Numer. Anal., to appear.

B. Lindberg, "Error estimation and iterative improvement for discretization algorithms," BIT, to appear.

R. D. Skeel, "A theoretical framework for proving accuracy results for deferred corrections," tech. rpt. UIUCDCS-F-80-892 (January 1980), submitted to SIAM J. Numer. Anal.

R. D. Skeel, "Ten ways to estimate global error," in preparation for SIAM J. Numer. Anal.

R. D. Skeel, "Equivalent forms of variable step multistep formulas," in preparation for Math. Comp.

J. R. Van Rosendale and R. D. Skeel, "Rapid solution of finite element equations on locally refined grids by multi-level methods," in preparation for Math. Comp.

J. R. Van Rosendale and R. D. Skeel, "Approximation of finite element equations on locally refined multi-level grids," in preparation for Math. Comp.

R. D. Skeel, "The order of accuracy for a deferred corrections algorithm," in preparation for SIAM J. Numer. Anal.

PERSONNEL

Principal Investigator

Robert D. Skeel, Assistant Professor

Other Faculty

Bengt Lindberg, Visiting Assistant Professor, 1975-76

Research Assistants

John R. Van Rosendale, 1975-79, Ph.D. in 1980, thesis entitled "Rapid solution of finite element equations on locally refined grids by multi-level methods"

Blake Ortman, 1977-78

Mary Ann Dahlquist-Downs, 1978-79

Thu V. Vu, 1979-80

INTERACTIONS

Spoken Papers

R. D. Skeel, "Scaling systems of linear equations," at the SIAM 1977 Fall Meeting, Albuquerque, October 31-November 2, 1977. Abstract in SIAM Review 20, 634.

R. D. Skeel, "Stiffly stable linear multistep formulas," at the ODE Workshop, Albuquerque, November 4-5, 1977. Abstract in SIGNUM Newsletter 13, 2, 8.